

Testing the Limits:
The Purposes and Effects of Additional, External Elementary Mathematics Assessment

By

Copyright 2011
Karen Ann Lombardi

Submitted to the graduate degree program in the Department of Educational Leadership and Policy Studies and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dr. John L. Rury (Major Advisor)

Dr. Mickey Imber

Dr. Jennifer Ng

Dr. Argun Saatcioglu

Dr. Kelli Thomas (Minor Advisor)

Date Defended: December 13, 2010

The Dissertation Committee for Karen Ann Lombardi
certifies that this is the approved version of the following dissertation:

Testing the Limits:
The Purposes and Effects of Additional, External Elementary Mathematics Assessment

Dr. John L. Rury (Major Advisor)

Dr. Mickey Imber

Dr. Jennifer Ng

Dr. Argun Saatcioglu

Dr. Kelli Thomas (Minor Advisor)

Date approved: December 13, 2010

Abstract

This mixed-methods case study focuses on the third through fifth grade classrooms at a public elementary school in a Midwestern urban school district where the Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) assessment is being implemented. According to the school district, the goals of these tests are: to show student growth in meeting the state benchmarks, to identify student strengths and weaknesses, and to provide teachers with information to direct content and pedagogy. Unofficially, another goal seems to be to hold classroom teachers accountable for the progress of their students by providing them with data on student concept attainment and measuring student improvement over the year.

Since the district and the NWEA stress the importance of utilizing these test results at the classroom level, the purpose of this research is to identify how MAP testing and the application of the test results are (or are not) utilized to inform mathematics instruction at the classroom level and to explore the effects of additional, external testing, particularly the effects on the teachers and students. Additionally, while the schools in the district do not have the option to exclude themselves from participation in this mandated testing program, a summative question as to the actual value of (or need for) the information provided by MAP testing will also be explored. Based on the results of this study, a more generalized understanding of the effects of additional, external standardized tests is also discussed.

This study, in particular, adds a unique perspective to the current literature on external testing because, unlike previous studies that have focused on state-mandated standardized tests, this study focuses on a district-mandated test that is used in addition to the state-mandated testing. Along with the current body of literature on assessment, this study can inform educational policies that challenge the culture of testing – the culture of data, pressure, and confusion – that is present in classrooms and schools throughout the country.

For
Sr. Elena Morcelli, AVI

Thank you for making me finish this and, most especially,
for encouraging me, praying for me, and putting up with me in the process.

Table of Contents

Acceptance Page	ii
Abstract	iii
Dedication	iv
Table of Contents	v
Chapter One: Introduction	1
Definition of the Problem	
Methodology	
Review of the Literature	
Expected Outcomes and Relevance	
Chapter Two: Why Are These Additional Assessments Being Used?	26
Data	
Methods	
Results	
Analysis	
Chapter Three: What are the Effects of This Additional Testing?	55
Data	
Methods	
Results	
Analysis	
Chapter Four: Is This Additional Testing Providing New Information to Teachers?	86
Data	
Methods	
Results	
Analysis	
Chapter Five: Education Policy and Additional, External Standardized Tests	111
Summary of Findings	
Strengths and Limitations	
Recommendations	
Shifting the Paradigm of Education Policy	
Concluding Thoughts	
References	136
Appendices	146

Chapter One

Introduction

I began teaching in urban public schools in 2003, when No Child Left Behind (NCLB) jargon was first entering the daily life of classroom teachers. School districts were struggling to make sure that teachers were highly qualified, that students made adequate yearly progress (AYP), and that no school would be labeled as failing. We were concerned about testing and funding and, therefore, about making sure that everything looked good on paper. This was a very discouraging way to start my career in teaching. Everyone was confused as to what these mandates really meant and would mean to our teaching and, yet, we were constantly aware of the pressure from these mandates in our daily work.

As a mathematics teacher, the pressures regarding the state tests were exceptionally high, since mathematics was one of the two subject areas (along with reading) to be tested yearly and used to determine AYP. Most department meetings involved matching state standards to examples of assessment questions. We analyzed data from district tests to find the weaknesses in our curriculum. We were asked to write tests first and then design our lessons around the assessments; in that way, everything that we taught was something important enough to be tested, and everything that we tested was something that the state might actually assess. We formatted at least some of our classwork, homework, and tests in the same way that the state would present questions. We completed practice tests, discussed test-taking strategies, and stressed to the students the importance of taking these tests seriously.

After three years in two different districts teaching secondary mathematics, I felt powerless to truly teach in a meaningful way within this system. So, I decided that I wanted to be involved with changing the system, and I chose a graduate program in education to focus on

education policy. I knew that I wanted my research to focus on the policy issues that most affected urban schools – those that struggled the most to find and retain highly qualified teachers, to meet AYP, and to successfully educate students (whether or not a policy might define the school as “failing”).

As I began my graduate studies, I worked as a graduate research assistant for three years on a research project under the Collaborative Evaluation Communities (CEC) National Science Foundation grant. Through this grant, I worked together with elementary teachers and administrators at several schools in a Midwestern urban school district to evaluate different aspects of their mathematics and science instruction. The school that I selected for this case study was one of the schools that seemed most open to our project’s involvement in their school. Since this school initially chose to focus on mathematics instruction, I spent most of my time during the 2008-2009 school year on the CEC project working with the teachers and administrators there. The teachers were enthusiastic about working on the grant, collaborating with each other, and trying new ideas in their classrooms. For example, this group of teachers decided to work through a modified lesson study process where they collaborated on planning a lesson. Then, they observed the lesson being implemented by one teacher in the grade-level, gathered observational data and student work samples, discussed what was effective and where the lesson could be improved, and then re-implemented the improved lesson in the other classrooms. Again, they gathered more data as they retaught the lesson in order to analyze and further refine the lesson for future use. Through this process, most of the teachers became very comfortable with me observing and helping in their classrooms. Over time, the teachers also became comfortable showing and discussing their struggles with their mathematics instruction. They knew that the information they shared would not be used “against” them or reported to

their administrators. The teachers also knew my background as a teacher – that I did not just dismiss their concerns or complaints but could actually empathize with them.

During my time in the schools involved in the CEC project, I saw constant reminders of the pressures of assessment and the requirements of NCLB permeating these schools, particularly at the classroom level. My first year on the CEC grant coincided with the first year that the school district piloted the Northwest Evaluation Association (NWEA) – Measures of Academic Progress (MAP) assessment. I saw the classroom disruptions from students missing instruction to take the tests and from dealing with computer problems. I heard the teachers complain about data received from these tests, including their confusion with the Rasch Units (RIT) scoring on these tests. And yet I heard administrators speak hopefully about what this new test would provide to their teachers in order to help better prepare the students for the official state assessment. During my second year as a research assistant, the district adopted MAP for all of its schools (which is when the school that was selected for this case study began using the test).

From my own experiences as a teacher, I already questioned the need for and benefits from all of these assessments. And, while working with these schools, I began to wonder if the teachers were receiving useful information from these additional tests that the district was now purchasing. This question became the initial inspiration for this study. While much research has and is being done on the effects of state assessment efforts, more localized policy decisions regarding assessment are not getting as much attention and, yet, are affecting the daily lives of teachers and students at least as much as the statewide tests.

There tends to be a bias in education policy to focus on the large-scale policies and the large-scale effects. My personal bias is to focus on the localized effects of school-level and

district-level policies (which are, of course, influenced by state and national policies). I believe that the story of what is happening in the nation's schools can only be fully understood at the level of implementation. I bring to my work a concern for the well-being of the individuals involved at that level – the students and the teachers. Like many educators, I am concerned that NCLB policies are doing more to improve test scores than to improve learning. However, I am even more concerned that policy makers are only concerned with outcomes (even meaningful outcomes) and are missing the problems with the means that are being used to arrive at those outcomes. What is being sacrificed to achieve these scores? What new obstacles and frustrations are inadvertently introduced into the process of teaching and learning when the data are collected? What important pieces of the classroom experience are being neglected by reducing complex educational outcomes to numbers? My research looks at the localized means being employed to attempt to meet policy mandates. While the MAP assessment is a test that is marketed nationally and is a clear response to the assessment movement that was revitalized by NCLB, it is voluntarily adopted by individual districts and primarily utilized to make school and classroom-level decisions. Therefore, the MAP assessment is an ideal example of one local initiative shaping the classroom teaching and learning experience in response to national policy.

Definition of the Problem

Defining Key Acronyms: NWEA, MAP, and RIT

Before discussing the specifics of this study, a familiarity with the NWEA, MAP testing, and RIT scores is helpful. NWEA provides an immense amount of information (including its own studies and reports) about its work, its assessments, and its data; and most of this information is readily available on the NWEA website (www.nwea.org). Rather than repeat information that is widely available to those interested, this section will provide a brief, general

overview of the organization responsible for developing the test (NWEA), the test itself (MAP), and the primary form of data the test provides (RIT scores).

Northwest Evaluation Association (NWEA). NWEA's primary purpose is the development and distribution of assessments. Corresponding with the assessments, classroom resources (such as DesCartes) and analytical tools as well as professional development are provided by NWEA. While the NWEA first began administering assessments in Portland, Oregon, in 1978, the MAP test is a more recent version of its assessment and one that has been the most widely used. MAP was first implemented in 2000, and, according to the NWEA website, by 2003, more than 1200 school districts and education agencies nationwide were using MAP. As of 2006, MAP testing had been adopted by over 150 public school districts in Kansas; additionally, private schools, individual public schools, and other educational entities within the state were also implementing the test (Northwest Evaluation Association, 2006a).

Measures of Academic Progress (MAP). MAP is a computer-based standardized test. It is adaptive in that the level of questions presented is adjusted based on the students' demonstrated abilities on previous questions. MAP is intended to be a formative assessment that identifies a student's instructional level. NWEA highlights the value of MAP over other standardized assessments because it is intended to show growth over time during and between school years and grade-levels.

Most often, schools (including the one chosen for this study) use MAP assessments in reading and mathematics, although NWEA has also developed assessments for science and language usage. This study focuses only on the use of the mathematics portion of MAP in third, fourth, and fifth grade classrooms. (Prior to grade-three, there is a different version of the MAP test, called MAP for Primary Grades; the format of this test is different, with fewer items in the

form of multiple choice responses, more graphics, and an audio component that reads questions aloud to students.) For the upper grade levels (third-grade and higher), the same MAP assessment is administered, meaning the same bank of questions is available to each student and items are selected for the individual student based on his or her previous answers (and his or her previous MAP results). The test consists of fifty-two multiple choice items; students must read the questions presented to them on their own and respond to each question before moving on. They cannot return to items that they have already completed. There is no time-limit on the test and students can save a test part-way through and return to finish testing later.

Rasch Units (RIT). The RIT scale is an equal-interval scale designed in such a way that every item on the test corresponds to a specific difficulty value. According to NWEA, “the scales were developed independent of grade level structure and therefore do not rely on student grade level for their meaning” (NWEA, 2008c, p.1). In other words, a score of 190 has the same meaning and indicates the same instructional level regardless of the students’ grade-level. As Traub and Wolfe (1981) explain, “Easy items discriminate among poor students to the same degree that hard items discriminate among good students; items from different subdivisions of the content domain and items in different formats discriminate equally” (p. 388). Because of the equal intervals and the independence of the score’s meaning from a specific grade-level that are both inherent to Rasch Unit design, scores can be meaningfully compared across grade-levels and can be averaged within a classroom, school, or district. Another key feature of Rasch models is that “the item difficulty defines the point on the ability scale where the probability of correct response for persons of that ability is .5” (Baker, 2001, p. 22). Due to this design, all students will answer about half of the questions presented to them correctly.

While there are few school-based assessments using RIT scales, this is a frequently-used tool in psychometrics. NWEA is not the only organization applying the Rasch Model to widely-used educational assessments; for example, the Third International Mathematics and Science Study uses the Rasch measurement. Bond and Fox (2001) call the Rasch model the “one readily accessible to . . . construct objective, additive scales . . . This model can help transform raw data from the human sciences into abstract equal-interval scales” (p.7). However, there are ways that the Rasch model falls short or, at least, should be used with caution both statistically (Traub & Wolfe, 1981) and in practice (Gipps & Murphy, 1994).

Purpose of this Study

The purpose of the research presented in this study is to identify the effects of additional external assessment on mathematics instruction at the classroom level. This study will begin by examining the question of why additional external assessments, specifically the Northwest Evaluation Association (NWEA) – Measures of Academic Progress (MAP), are being used. Next, the study will explore how this testing affects teachers, students, and overall classroom instruction. Then, the study will address the question of whether the data from these assessments equip teachers with information about their students that they do not already have from other assessment efforts. Based on the answers to these questions, this research will offer a recommendation regarding the value of continuing to use additional external assessments.

This case study will focus on third through fifth grade classrooms at a public elementary school in a Midwestern urban school district. During the 2006-2007 school year, the district decided to adopt the MAP test as a mandated test for all schools in the district. The test is administered at every grade level two to three times per year to measure progress in mathematics (and, for many grades, in reading, as well); it takes the place of local assessments that were

previously administered in both subjects (School District Reference, 2006c). An advantage to this test over previous assessments is that it is conducted on the computer so that teachers have the results within twenty-four hours.

The school selected for this study participates in numerous mathematics assessments. In addition to the yearly state assessment tests, the school uses an online test-builder created by the state to develop mathematics assessments for each unit. The results from these assessments are provided to each classroom teacher so that he or she can easily identify the standards and types of questions with which students struggled the most and can reteach or readdress these standards in future lessons. Additionally, the teachers use the daily assessments, class activities, and homework assignments for continuous student assessment, and some teachers use the curriculum's provided unit assessments.

On top of all of these data, teachers are then given the very dense results from the MAP testing; the stacks of results they get back from these tests are filled with columns of numbers for each student, including many scores and score ranges in forms unique to MAP testing and, therefore, somewhat complicated to read and analyze. There is concern that these results are, therefore, not being used effectively. In addition, with the three-times per year that this test is implemented at the school selected for this study, students are frequently missing other, regular instructional activities. According to notes from a summer meeting on MAP testing in the school district, each subject test usually takes the student about 90 minutes, although there is not a time limit (School District Reference, 2006a). The students, many of whom are identified as English language learners, are often frustrated by the test despite the fact that the MAP test adapts the level of the questions to each student as they take it. Decreasing the difficulty of a mathematics concept does not, generally, decrease language barriers for limited English proficiency students.

Since the district and NWEA stress the importance of utilizing these test results at the classroom level, the purpose of this research is to identify how MAP testing and the application of its results are (or are not) utilized to inform mathematics instruction at the classroom level as well as how this tool could become more useful to classroom teachers in the future.

Additionally, while the schools in the district do not have the option to exclude themselves from participation in this mandated testing program, a summative question as to the actual value of (or need for) the information provided by MAP testing will also be explored.

Methodology

The initial part of this study involved background information on the implementation of MAP testing. Contextual information for this school and school district came from analyzing the district webpage and the district's internal website, as well as school and district archives of training materials, memos, and other related materials. Specific information on MAP testing was gathered through some of these same sources but came, primarily, through the NWEA website and reports. These included documents available in the public domain of the NWEA website as well as those reports to which only subscribing districts have access.

School Demographics

The school chosen for this case study is an urban elementary (K-5) school in the Midwest. Eighty-four percent of the students are economically disadvantaged, qualifying for free and reduced lunch. The student population is racially and ethnically diverse, with fifty-six percent identifying as African American, nineteen percent as White, and seventeen percent as Hispanic; about eight percent of the student population represent other racial/ethnic backgrounds. Twenty-percent of the student population fall under the category of English language learners,

although even more students than this would not claim English to be their first language and, therefore, also need language support in the classroom.

Participants

The primary participants in this study are the eleven third, fourth, and fifth grade teachers at this single urban school. Three school administrators (the two instructional coordinators and the principal) also participated in semi-structured interviews. All of these educators were already involved with a current research project under the Collaborative Evaluation Communities (CEC) National Science Foundation grant. As a graduate research assistant on the CEC grant, I already had access to this school, these teachers, and the student data.

The teachers' years of teaching experience ranged from one to thirty-two years, with just under half teaching for five years or fewer. The longest that a teacher had been at this one school was nine years. All of the teachers hold full licensure to teach elementary grades with some also holding licensure for older grades or holding endorsements.

Instruments

The only specific instrumentation created for this study was the teacher surveys – where the teachers predicted student outcomes on the MAP test, demonstrated their ability to interpret RIT (Rasch unit) data, and responded to statements about their view and use of these assessments. The prediction portion of the survey included a row for each student, a column with his or her Fall RIT scores, and an item to predict his or her Spring RIT scores, along with four columns to predict ranges (low, average, or high) for each student in the four sub-categories measured on the mathematics assessment (Algebra, Data, Geometry, and Number & Computation).

The survey instrument was based on several studies (Green, & Stager, 1986; Moore & Waltman, 2007; Pedulla, et al., 2003; Sever, 2004) as well as on the perceived needs and potential contributions of this school and the participating teachers. Analysis of the prediction data followed the models presented in previous research, as well (Miller & Davis, 1992; Miller, S & Mee, 1991; Pezdek, Berry, Renno, 2002).

Additionally, I developed semi-structured interview protocol for both the teacher interviews and administrator interviews. Each interview was digitally recorded and then transcribed. I also kept detailed notes during all observations of classroom lessons and teacher meetings.

Mixed Methods

This study is intentionally designed as a mixed methods case study. Tuhiwai Smith (1999) states that all “research is about satisfying a need to know” (p. 170). In qualitative research, the purposes are “to *explore, explain, or describe* a phenomenon. Synonyms for these terms could include *understand, develop, or discover*” (Marshall and Rossman, 2006, p. 33). Understanding is also the primary purpose of mixed methods research; the goal is to gain a greater and richer understanding of problems, questions, and phenomena than could be obtained through a qualitative or quantitative approach alone (Creswell & Plano Clark, 2007; Greene, 2007). While potentially useful (and used) in other fields, Greene (2007) sees mixed methods as an important approach in social inquiry because this form of inquiry is so complex and contextual. The *Journal of Mixed Methods Research* adopts a definition of mixed methods research that seems similar to Greene’s:

. . . [M]ixed methods research is defined as research in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or program of inquiry (Teddle & Tshakkori, 2006, p. 15).

Mixed methods research is used for a variety of purposes, but most often it is chosen for the purpose of complementarity; when the purpose of the mixed methods design is for complementarity, “methods are intentionally chosen or designed to measure different facets of the same complex phenomenon” (Greene, 2007, p. 101); similarly, but less frequently, the purpose is expansion where “different methods are used to assess *different phenomena*” (p. 103). In a sense, several separate studies are conducted in order to, in the end, have a fuller understanding of a larger situation. Together, these studies may encompass many different methods and methodologies in order to explain the overall school climate. However, each individual study may have used a more traditional, distinct methodology (and the appropriate methods for that methodology).

In some ways, I see my approach to mixed methods research as a qualitative approach that includes the possibility of using quantitative methods. This study, in particular, falls within the definition of a complementarity stance to qualitative research. However, some of this research could be framed to push it closer to a role of expansion. Creswell (2003) made an interesting, practical point for novice researchers when he said, “[H]aving a major form of data collection and analysis and a minor form is well suited for studies undertaken by graduate students” (p. 212). So, while my ideal may be to approach mixed methods like Greene (2007), this advice from Creswell is guiding my initial work with mixed methods. I see myself having qualitative data as my “major form of data collection” and quantitative as my “minor form.”

Types of Data

In many ways, the concept of mixed methods seems simple. One’s research uses qualitative data and quantitative data to gain a more thorough understanding of a topic of interest. Additionally, I do not want to limit myself or my research by one label – qualitative or

quantitative. Practically speaking, the mixed method approach is most appropriate in this study because there are several questions that need to be addressed that could not be adequately answered with only one measure. Some flexibility was required to adapt as new goals, issues, or even new questions emerged.

Qualitative data included classroom observations, teacher interviews, teacher surveys, administrator interviews, and available district documents. Observing in the classroom provided an opportunity to see, specifically, how the MAP testing is conducted, the regular mathematics instruction in the classrooms, and any factors that could go unnoticed or unexamined otherwise. To supplement these observations, teacher interviews and surveys allowed those actually responsible for the test implementation and the use of data to explain the effects of MAP testing in their classrooms. Administrator interviews and district documents helped confirm the goals of the MAP testing and the expectations of how MAP should be implemented, how the data are intended to be used at the classroom level, and any other unwritten goals or perceptions of the program.

This study also included several forms of quantitative data. Prior to the Spring administration of the MAP assessment, the teachers completed a paper survey where they predicted an overall score for each of their students as well as a score range in the various sub-categories measured on the MAP mathematics assessment. These predictions were then compared to the actual student results on the Spring MAP assessment to determine if the test itself provided information to the teacher that she would not have otherwise known. True and false questions on the survey that were intended to provide data on the teachers' level of ability to understand and interpret RIT scores were also examined quantitatively through the use of descriptive statistics.

Data Collection

An important strength of this work is that it is that much of the data were obtained directly from the study participants through surveys, interviews, and observations. Similarly, additional data came from utilizing existing instruments that were already in use at the school and reviewing existing documentation at the school or through the district website. The additional time commitment on behalf of teachers to complete the survey and prediction instruments was worked into their Wednesday afternoon professional development activities. Teachers also generously shared some of their prep-periods (their limited time during the school day without students), to participate in the interviews. The three administrators also made time in their tight schedules to meet for individual interviews during the school day.

While preliminary data, particularly district documents regarding MAP testing, were collected since my work began with the CEC project in 2006, most of the data collected specifically for this study were obtained in the spring of 2009. The teachers completed the surveys in April 2009. Then, throughout April and May 2009, data collection activities included classroom observations, observations of MAP testing, and teacher and administrator interviews. Student test results were available through MAP testing almost immediately after the students completed the tests in May.

Potential Limitations

A potential limitation of this study was the political viability. The teachers needed to have enough buy-in to willingly allow classroom observations, to participate in interviews, and to complete the survey. However, since I already had an established relationship with this group of teachers through the Collaborative Evaluation Communities (CEC) project, this did not seem to be much of a struggle with the majority of teachers.

Possibly a greater political liability, however, came from outside the school itself. The district has invested in the MAP program and, therefore, has an interest in the demonstration of its success. At the school and district level, the administrators were under pressure to produce improved results in student mathematics achievement and to support district efforts to improve mathematics instruction. Therefore, the information provided in the interviews, especially by the administrators, may not fully represent their personal views but, rather, might parrot the standard district response to the questions.

Review of the Literature

Issues Surrounding Assessment and Accountability

The idea of accountability and testing is not new, having gained prominence with the publication of *A Nation at Risk* in the early 1980s. Guthrie and Springer (2004) note that *A Nation at Risk* was a catalyst for the increased federalization of education policy as well as the greater focus on test scores as the primary measure of student achievement. While the standards movement has been prominent in American education and, specifically, in mathematics education for over two decades, it has had its greatest effects since the mandates of No Child Left Behind (NCLB) were established in 2001. NCLB ushered in a greater focus on accountability, particularly through high-stakes standardized testing.

The term “high-stakes testing” refers to assessment programs that produce results that are used to determine rewards and sanctions at the institutional level, at the student level, or oftentimes at both levels (Pedulla et al., 2003). For students, these high-stakes may include successfully completing or being retained in a specific grade, passing or failing a course, or determining whether or not they qualify to graduate (Green & Stager, 1986; Smith, 1991).

Institutional level consequences may include changes in district funding, district management, teacher pay, or even teacher and administrator job security (Smith, 1991, Pedulla et al., 2003).

While assessment can support learning, particularly when viewed as formative, the purpose of assessment in the United States is often summative (William, 2007). Under NCLB, these summative assessments also become the sole source of data for punitive measures toward schools, making them “high stakes” tests. Ravitch (2010) specifically pinpoints the signing of NCLB as the point that “changed the nature of public schooling across the nation by making standardized test scores the primary measure of school quality” (p. 15). Because the stakes are high, these assessments often become the basis for determining the curricula and instruction (Cochran-Smith & Lytle, 2006; Gilman & Reynolds, 1991). Even prior to NCLB, however, researchers cautioned that “the increase in the amount of testing, as well as the misuse of test scores, increases tensions in education, to the detriment of children and teachers” (Haladyna, Hass, & Allison, 1998, p. 264). These concerns are magnified under the pressure of NCLB, as researchers explain that “high-stakes testing environments are related to instructional changes and test-preparation practices that may result in higher scores without actually increasing student achievement” (Moore & Waltman, 2007, p. i). Researchers also warn that an emphasis on test-preparation activities can actually invalidate or “corrupt” the accuracy and meaning of student test scores (Abrams, Pedulla, & Madaus, 2003; Koretz, 2008; Nichols & Berliner, 2007; Ravitch, 2010).

Formative Assessment

Educators understand the need for assessment. Good teachers are always doing informal, formative assessments of their students’ learning. And there is an obvious need for formal assessments, as well. As the National Research Council (2001) explains, “Information about

students is crucial to a teacher's ability to calibrate tasks and lessons to students' current understanding and skills" (p. 349). And, when used for this purpose, these assessments are formative and useful. However, large-scale assessments, especially those that are used as measures of external accountability, are often viewed by teachers as taking time away from classroom instruction and interrupting the flow of teaching and learning throughout the year (Gilman & Reynolds, 1991; Glaser & Silver, 1994; Mehrens, 2002; Smith 1991). Teachers also express concern about the stress these assessments put on the students in their classrooms (Mehrens, 2002; Smith, 1991).

Additionally, teachers are often overwhelmed with the amount of data received from these larger assessments and are not often given support in determining how to use these data to inform instruction. Darling-Hammond (2003) emphasizes that assessment and standards-based reform can only help to improve student learning if teachers *use* the data in their instructional decisions. However, Gardner (1982) explains that "there is substantial misunderstanding . . . among many educators of the meaning of tests scores" (p. 3). This misunderstanding can lead to detrimental misuses of test results (Gardner, 1982; Sever, 2004).

Black and Wiliam (1998) explain that most education reform initiatives "are not aimed at giving direct help and support to the work of teachers in classrooms" (p. 140); they cite the need for quality formative assessments that can be used by teachers to provide more effective instruction for their students. In essence, Black and Wiliam (1998) provide a narrower definition of formative assessment when they state, "For assessment to function formatively, the results have to be used to adjust teaching and learning; thus a significant aspect of any program will be the ways in which teachers make these adjustments" (p. 141). In other words, if teachers are not

making the adjustments based on the results of these assessments, then the assessments are not formative.

Classroom Level Effects of External Assessment

At the classroom level, external assessments directly affect students and teachers. Both groups have increased anxiety and stress due to these tests. In addition, the changes in instruction that can be attributed to more external assessment can hinder the teaching and learning process, in general. McNeil (2000) discusses the conflict that inherently exists between the schools' purposes of "processing aggregates of students through regularized requirements of the credentialing process" and the purposes of "nurturing individual children and equipping them with new knowledge and skills"; she goes on to explain that "at the point of the tension – where the two oppositional faces intersect – are the children, the teacher, and the curriculum" (McNeil, 2000, p. 11).

Students. Students often experience high levels of anxiety due to assessments, especially those that are external and high-stakes (Black & Wiliam, 1998; Gilman & Reynolds, 1991; Haladyna, Hass, & Allison, 1998; Wheelock, Bebell, & Haney, 2000). One study found that "more elementary and middle school teachers than high school teachers reported that their students are extremely anxious and are under intense pressure because of the state test" (Pedulla et al, 2003, p. 2). For high school students, high-stakes tests have been reported as a cause of dropping out of school (Abrams, Pedulla, & Madaus, 2003; McNeil, 2000). Teachers report increased anxiety in their students and a decrease in their confidence when external assessments have been added to their schools (Jones, et al., 1999). Wheelock, Bebell, & Haney (2000) directly studied the students themselves through an analysis of their drawings, finding that testing prompted drawings of students who were "anxious, angry, bored, pessimistic, or

withdrawn” (p. 1). Similarly, Stiggins (2004) describes some student reactions to high-stakes tests as “deflating, discouraging, and defeating” (p. 24). Barksdale-Ladd and Thomas (2000) also discuss the overwhelming stress placed on all involved, including the students, in environments of high-stakes testing.

Teachers. Research over the past two decades has indicated that required assessments cause unintended negative consequences for teachers. In a qualitative study of two schools in an urban district, Smith (1991) identified six main effects on teachers of external, mandated testing: (1) feelings of shame and anger based on the public nature of test results; (2) feelings of dissonance and alienation due to the perceived lack of relevance of the testing; (3) feelings of anxiety and guilt over the belief that this testing negatively affects young children; (4) reduction in available time for instruction; (5) limitation of teachers’ ability to adapt, create, or diverge by restricting the curriculum to the narrow range of tested material; and (6) use of fewer teaching methods and perception of teacher work being devalued. Similar results have consistently been found in more recent literature on the effects of high-stakes testing on teachers (Au, 2007; Black & Wiliam, 1998; Cochran-Smith & Lytle, 2006; Lipman, 2004; McNeil, 2000; Watanabe, 2007).

Lack of instructional time is often cited as an unintended consequence of increased testing (Abrams, Pedulla, Madaus, 2003; McNeil, 2000; Pedulla et al., 2003; Smith, 1991). The testing itself takes time away from instruction, as does test-prep activities. Abrams, Pedulla, and Madaus (2003) point out that “not only do teachers . . . report that they are spending more time on tested content, but state tests, especially those with high-stakes attached, are also influencing the frequency and manner in which teachers assess their students” (p. 24). Potentially even more detrimental to instruction throughout the year, high-stakes testing often narrows the material taught to that which is emphasized on the test; this reduces the time spent on non-tested material,

non-tested subjects, and non-essential activities (Pedulla et al., 2003; Rothstein, Jacobsen, & Wilder, 2008). Mandated testing can also lead to educational methods and activities that do not align with the teachers' personal ethics related to good practice (Lai & Waltman, 2008).

Recent research, conducted in response to the high-stakes testing of NCLB, cautions that the pressure felt by the teachers is mostly related to a focus on increasing test scores rather than on improving student learning (Moore & Waltman, 2007). Elementary teachers tend to feel more pressure than middle and high school teachers (Moore & Waltman, 2007; Pedulla et al., 2003). Moore and Waltman reported that this pressure (which teachers perceived to come primarily from administrators and the government) was strongly associated with teachers reporting a test-centered focus, instructional changes, questionable or potentially unethical test preparation practices (which are examined in more detail by Lai and Waltman, 2008), and decreased morale. This decreased morale is also associated with a desire of teachers to transfer out of tested grade levels (Pedulla et al., 2003).

Instruction. The effects of assessment on instruction are of particular concern in that they directly affect both students and teachers. Since instruction is foundational to schooling's purpose and function, if testing negatively influences the quality of instruction, then it directly undermines the basic goals of education. Pedulla et al. (2003) report that "a substantial majority of teachers at each grade level indicated that state testing programs have led them to teach in ways that contradict their ideas of sound instructional practices; this view was particularly pronounced among elementary teachers" (p. 3). Diamond (2007) found little evidence of changes to pedagogy but strong evidence in changes of content in response to testing policies, including the narrowing of the curriculum. Diamond argues that education policy, therefore, does not have as strong of an effect on instruction as it is often reported to have. However, he

fails to grasp both the strong impact of a narrowed curriculum and the possibility that the lack of change in pedagogy, itself, is another factor that can be attributed to increased testing; while Diamond finds little change in pedagogy prior to and after the addition of more external testing, the testing policies often provide teachers with more incentive to *not* change their pedagogy as the teachers feel their traditional instruction (or the didactic instruction that they see other teachers using, as Diamond discusses) is “safer” than trying newer, alternative modes of instruction.

These concerns regarding instruction are repeated even more frequently in urban schools and schools with high-minority populations. The increased use of standardized assessments leads to a standardization effect that “reduces the quality and quantity of what is taught and learned in schools,” creating “inequities, widening the gap between the quality of education for poor and minority youth and that of more privileged students” (McNeil, 2000). Underserved populations often have schools that “give more attention to testing in planning and delivering their instruction, spend more class time in test preparation, . . . report more school attention to testing, and are likely to give less attention to nontested content and thinking skills” (Herman, Abedi, & Golan, 1994, p. 481). Similarly, Glaser and Silver (1994) cite the tendency in high-minority urban classrooms for teachers to consistently present skills in the same form as is found on the assessments.

In an effort to improve instruction, there has been a movement in education toward “data-driven” instruction. In the past, teachers have tended to ignore assessment data because of their lack of understanding of the data and/or the perceived lack of relevance (Sever, 2004). Proponents of data-driven instruction, however, do emphasize that standardized tests should only

be *one* measure (rather than the only measure) used in making instructional decisions (Blink, 2007; Sever, 2004).

Predicting Achievement Scores

The main purpose of school assessment is to measure student knowledge and performance. The underlying assumption seems to be that formal assessments provide more accurate and meaningful information than could otherwise be ascertained. In particular, teachers are often criticized as poor judges of student abilities due to bias or other errors. Therefore, “[i]t is commonly argued that commercial tests provide teachers with valuable information about the abilities and deficiencies of their students, from which it follows that teachers who rate their students without such information will often be in error” (Egan & Archer, 1985, p. 25). But do all of the assessments used in classrooms really provide information that students, parents, and teachers would not otherwise know?

A few studies from the 1980s specifically examined teacher judgments and predictions for student performance on standardized tests and found that teachers’ predictions were reasonable and relatively accurate (Coladarci, 1986; Egan & Archer, 1985; Hoge & Butcher, 1984; Leinhardt, 1983; Wright & Wiese, 1988). Doherty and Conolly (1985) reported that teachers generally overestimated student performance in mathematics and English. While Doherty and Conolly were able to identify academic competence as the greatest influence over accurate teacher predictions, they also determined that the teachers in the study did a better job predicting the scores of students who reported having a positive view of their relationship with their teacher. Hoge and Butcher (1984) reported high correlations between teacher judgments and actual achievement test scores overall, but they also found that the level of accuracy in judgment varied according to student ability level, where teachers overestimated the scores of

students of high-ability but underestimated the scores of students with low-ability. This finding was also supported by Coladarci (1986). Additionally, Coladarci found variations in accuracy by teacher, with some teachers having greater variability in accuracy of predictions than others; this teacher effect was significant for mathematics concepts, while it was not found to be significant for mathematics computation, reading vocabulary, or reading comprehension. A similar effect between classrooms was also reported by Helmke and Schrader (1987).

Since the 1980s, few studies have appeared in the literature that involve teacher prediction of student test scores. Instead, parental predictions have, in a few cases, been the focus for studies. Pezdek, Berry, and Renno (2002) looked at how accurately parents could predict their elementary-aged child's performance on an assessment. They found that parents generally overestimated their children's mathematics assessment scores. This finding is consistent with previous studies examining parental prediction of children's performance (Miller & Davis, 1992; Miller, Manhal, & Mee, 1991) where parents again predicted scores significantly above their child's actual level of performance. Miller, Manhal, and Mee (1991) had mothers and fathers predict their second-grade and fifth-grade children's performance on an assessment of cognitive ability; both mothers and fathers similarly overestimated their *own child's* performance but more accurately predicted the *average* performance of children at each grade level. Miller and Davis (1992) specifically looked at the predictions of mothers and, again, found a pattern of overestimation. However, they also compared teacher predictions, peer predictions, and self predictions to actual performance, finding that the teachers' predictions were just as accurate as the mothers' predictions and the teachers' predictions were also more highly correlated with student performance than the mothers' predictions. Peer and self predictions were both less accurate and not highly correlated.

Expected Outcomes and Relevance

Particularly in regards to reforming No Child Left Behind, more research is needed on issues of assessment and accountability, including the effects of current (or future) policies on classroom level instruction. The focus on changing test scores provides an incomplete and often inaccurate picture (Hiebert, 2003). “Because of the important social and political impact of high-stakes testing in mathematics, further research is urgently needed. . . [T]he very fact that the stakes are high should underscore the critical nature of this research agenda” (Wilson, 2007).

Additionally, much of the research on the effects of assessment comes from surveys of teachers and administrators (Mehrens, 2003). Therefore, Mehrens cites the need for more classroom observation data in this area. Similarly, Smith (1991) explains:

To understand fully the consequences of high-stakes external testing on teachers, one must look beyond their verbal statement to underlying meanings within the institution. One must sit with them in the . . . faculty meetings . . . , observe their everyday classroom life throughout the school year, watch their sometimes frenzied preparation for the tests themselves, examine what topics and subject matter gets slighted or left by the wayside for the sake of the test, and finally learn what reactions to these experiences are incorporated into the teachers’ . . . definitions of teaching (p. 8).

McNeil (2000) also identifies “serious gaps” in current research, including “the absence of critical scholarship that . . . builds theory from what goes on inside schools” (p. 7-8). Therefore, this study contributes to the current literature by providing a more complete view of the classroom level effects of additional standardized assessments through observational data triangulated with the teacher surveys, teacher and administrator interviews, and document analyses.

According to the categories outlined by Hoge and Coladarci (1989), this study includes a “direct evaluation of teacher judgment” whereby the teachers predict student outcomes on a specific achievement test. The judgment measures include a modified or enhanced version of

what Hoge and Coladarci classify as a *grade equivalence* measure. However, unlike the other studies reviewed by Hoge and Coladarci that used grade equivalence measures, the MAP test allows for *peer-independent judgments* rather than being norm-referenced. This means that teacher predictions for the individual students' scores are independent of one another since students are not compared to each other in the determination of MAP scores. Data collected on predictions and results in this study could, therefore, be examined at the student-level as well as the classroom, grade, and school levels.

This study is also different from previous studies in that it is not focused on a state-mandated standardized test but, instead, on a district-mandated test that is used in addition to the state-mandated testing. Therefore, based on the data collected in this study, I am able to provide informed recommendations to improve assessment (amount, forms, use) so that unnecessary testing can be eliminated and necessary testing can be utilized to its fullest potential.

In their review of seventeen studies that looked directly or indirectly at teacher judgment of student achievement, Hoge and Coladarci (1989) conclude that teachers' judgments are relatively accurate and that the research, therefore, "does not support the total rejection of teacher judgments that one sometimes encounters" (p. 309), and they further assert that "it is time that we began giving [assessment of achievement provided by teachers] the same attention accorded other types of measuring instruments" (p. 310). Egan and Archer (1985) state: "We can no longer assume that teachers find tests helpful because they tell them something they did not know before" (p. 33). With these assumptions in mind, this study explores the usefulness and influence of the MAP testing at both the classroom and district levels.

Chapter Two

Why Are These Additional Assessments Being Used?

The Northwest Evaluation Association markets its products to states and school districts throughout the country. In some states, such as Colorado, it has been approved by the Department of Education as a provider of formative assessments (Northwest Evaluation Association, 2009b). In other states, such as Kansas, it has been successful at marketing its Measures of Academic Progress (MAP) assessment at the district-level, where individual school districts decide if this test is a worthwhile investment for their purposes. This study focuses on the case of one school in an urban district in Kansas. Therefore, the district purposes of the assessment are a key component to understanding the use of the MAP assessment in the school. Along with the official statements of the school district, as analyzed in district documents on the implementation of the MAP testing, the administrator and teacher perceptions of the purposes of MAP and how these purposes are being fulfilled in practice are of interest to this research.

In this chapter, the perceived purposes of the MAP assessments will be explored alongside the practical application of these tests at the classroom level. The data presented will address what administrators as well as teachers believe is the purpose or value of these assessments. Beyond the purposes, however, this chapter will also explore how teachers perceive these tests and, in particular, how they perceive their value in practice. Together, these questions help address the larger question of the ways in which MAP testing is meeting or is failing to meet its intended purposes at the school and classroom levels.

Data

The purchase and use of the MAP assessment is a decision made at the district level. District administrators need to believe that there is a sufficient benefit from these tests to justify

the cost of purchasing and implementing them. Therefore, the first source for identifying the purpose of the MAP testing is district documents. Along with these documents, the three school-level administrator interviews enhance the understanding of the value the administration sees in requiring the use of these tests. However, the teachers' perspective on the value of MAP testing is often quite different from the administration. As revealed in the eleven teacher interviews, the teachers' views of the test, their discussion of its value and its drawbacks, and their use of these data often highlight a contrast with the stated district goals for MAP.

Quantitative data from the teacher surveys are also included to answer the questions addressed in this chapter. In particular, descriptive statistics amplify and clarify the views expressed in the teacher interviews regarding their views of MAP testing. A section of the teacher survey was also comprised of items from the Mathematics Teaching Efficacy Beliefs Instrument (MTEBI) (Enochs, Smith, & Huinker, 2000). These items were analyzed to compare the teachers' mathematics teaching efficacy and outcome expectancies, which helped explain some of the dichotomies or apparent contradictions present in the teacher responses.

Methods

Document Analysis

Documents from the school district website and NWEA website were collected through internet searches as well as through access to the intranet (the internal site that teachers and administrators can access) of the school district. In reviewing these documents, the primary focus was to find the stated goals of the MAP testing as well as to gain a better understanding of the uses and implementation of the test. The goals of the school district and of NWEA, as expressed through these documents, is presented in this chapter as the foundation for examining

and analyzing where teacher and administrator perceptions of MAP may harmonize or conflict with the official goals.

Administrator Interviews

Transcriptions and audio recordings of the three administrator interviews were reviewed repeatedly to develop overall themes or categories of responses. These categories included: purposes of the MAP, benefits of the MAP, concerns about the MAP, issues related to time for administering and using the MAP, uses of the MAP data, comparison of the MAP to the state test, effects of the MAP on teachers, effects of the MAP on students, and other related categories. Some of these categories came directly from the interview protocol (provided in Appendix A). For example, the administrators responded to the question: “How do you use the NWEA-MAP data?” Therefore, many of these responses were placed directly into the initial category of “uses of the MAP data.” Other categories, however, came from common themes introduced by the administrators that were not directly asked or anticipated in developing the interview questions. For example, even though the comparison of the MAP to the state test was not an intended area for exploration in the interviews, all of the administrators included this comparison in their interviews and, therefore, the area needed to be addressed in the analysis for this study.

Comments from each interviewee were organized into these broad categories – with each administrator’s text receiving a different color in the document in order to keep their comments distinct from one another. Once organized in this way, the text was coded in order to find the more specific themes that would emerge. For example, the administrators all spoke about identifying “strengths and weaknesses” from the MAP data. As these comments were coded and organized, it became clear that there were different ways in which they were using these terms – speaking of the strengths and weaknesses of individual students, of the students collectively, and

of instruction. These were organized into the categories now found in Table 2.1. In this case, the creation of the table reflects the process of the initial analysis, i.e. the coding and organizing, of these data. Other tables highlight more of the beginning stages of the interpretation of what had already been coded, such as in Table 2.2, where contradictions in comments about the alignment between MAP and KCA tests are compared to each other.

Teacher Interviews

The process of transcribing, categorizing, coding, and identifying emerging themes in the teacher interviews mirrored the process of analyzing the administrator interviews. Again, some broad categories came directly from the interview questions posed to the teachers (see Appendix B), while other categories emerged from the common responses brought up by teachers during the interviews that were never asked of them directly. For example, in cases where teachers were directly asked about a topic that became a broad category in the analysis, such as the category on the use of data, the responses were further coded and analyzed within this broad category to reveal the more subtle differences between *having* data and *using* data, which is presented in Table 2.3. Conversely, a category emerged about the teachers' desire for authenticity in standardized test scores; this was a very interesting and unexpected part of the data to explore, as summarized in Table 2.4, but none of the interview questions asked directly about this theme.

Survey Data

The data analyzed in this chapter also include ten items from a teacher survey. (A copy of the survey can be found in Appendix C.) The items from the survey that were used in this chapter relate to the teachers' perceptions and opinions of the NWEA-MAP assessment as well as twenty-one items from the MTEBI scale (Enochs, Smith, & Huinker, 2000) to measure

teachers' mathematics teaching efficacy and outcome expectancies. A four-point Likert scale was used on this survey for all of the items included in this portion of the study. The initial ten survey items presented in this chapter were adapted from a national survey used in a study by Pedulla, et al. (2003); their survey was given to classroom teachers in order "to collect information from those who witness the effect of state-mandated testing firsthand: classroom teachers" (p. 20). Only questions relevant to the purposes of this study were included. For example, questions on media coverage of testing were eliminated, as were questions on the format of state tests and computer access. Most of the questions included were only slightly revised from referring to state testing to, instead, asking about the NWEA-MAP assessment. The use of the MTEBI instrument, specifically, on a four-point Likert scale mirrored the use of this instrument by Akinsola (2008). The four-point Likert scale was chosen in order to allow for the instruments' questions to be used without distinguishing them from the rest of the survey format and, more importantly, by forcing teachers to take a stance by the elimination of a neutral or "uncertain" response option.

The MTEBI items were analyzed based on the previous use of the instrument in research (Akinsola, 2008; Enochs, Smith, & Huinker, 2000), finding an average "score" (out of 4) for each teacher on the two subscales: mathematics teaching efficacy and outcome expectancies. The mean score for the eleven teachers was then found (see Table 2.6). In this study, a matched pairs t-test was also conducted to see if the difference of these means was statistically significant.

The ten survey items regarding the teachers' views of the purposes and use of MAP were analyzed in terms of the percent of teachers in agreement with each statement. Teachers indicating that they agree or strongly agree were included in the statistic "percent agreeing" as shown in Table 2.5. Strong agreement was noted where it was included in one of the

percentages. The purpose of this analysis was descriptive, in order to reveal items with strong or weak overall teacher agreement.

Results

Qualitative Results: What do administrators believe is the purpose or value of these assessments?

According to a district newsletter for employees, the stated goals of the MAP assessment are: (1) to show student growth in meeting the state benchmarks; (2) to provide each student (and their parents) with a report showing strengths and weaknesses; (3) to provide the teachers with information to direct classroom content and teaching strategies; (4) to provide the teachers and students with expected achievement gains (School District Reference, 2006c). The Northwest Evaluation Association (NWEA), the creator and supplier of the MAP, believes that the results from the MAP assessment should be used, along with state assessment data and local classroom assessments, to triangulate student instructional levels and specific areas of strengths and weaknesses (NWEA, 2006).

In this particular school district, MAP testing is viewed as complementary to the state's computerized formative and summative tests. According to the Coordinator of Assessment at the district's office on educational research and assessment, "MAP locates the student instructionally and 'maps' how to get where [the computerized formative tests] tells us the student needs to be" (School District Reference, 2006d). With this focus, the results from MAP testing are intended to be used primarily by the classroom teacher to monitor individual student progress and the progress of the class as a whole. This is supported by a link on the school district's internal website, an internal bank of information for teachers, to the NWEA website's resource "10 Ways

to Use the Class Breakdown by Overall RIT and Class Breakdown by Goal Reports in the Classroom” (NWEA, 2005).

The school-level administrators interviewed for this study, however, reveal some purposes in MAP use that differ from the district’s stated goals. While many of the same official reasons resonated in these interviews including the formative value of monitoring student growth throughout the year, unofficially, another goal emerged in all three interviews: to hold classroom teachers accountable for the progress of their students by providing them with data on student concept attainment and by measuring student improvement over the year. In other words, administrators view these tests not solely as a measure of student strengths, weaknesses, and progress but also as a measure of teacher strengths and weaknesses. However, the MAP assessment is not designed to judge teacher quality nor is this an official reason that the district gives for adopting and continuing to use this test.

The principal of the school explained that one of the benefits of the MAP test is that it “tells the teacher a little about how she’s teaching.” At another point in the interview, the principal stated that “teachers have to know where their strengths and where their weaknesses are.” Beyond just self-awareness about their teaching, however, the principal uses the MAP results to form judgments about the teachers. When responding to a question on how she uses the MAP data that she receives, the principal explained that she meets with the two instructional coaches (the other two administrators interviewed at the school). Here is how she describes what they do: “We go over the data. We talk about, you know, where did we make our gains? Where were the teachers strong? Where were they not strong, you know, apparently in their instruction? We determine what it is that we need to give teachers in professional development that’s going to help them.”

This view of the test as a measure of teacher quality was also echoed by one of the instructional coaches, the one responsible for the fourth and fifth grades. She explained that “the teachers feel a greater accountability to use their instructional time well. That I appreciate.” She later made a similar point: “I think it increases teachers’ stress level because they do feel more accountable, and I see that as both a positive and a negative.” After a MAP assessment, one of her purposes in analyzing the data she receives is to “look at teacher average growth. In other words, which of the [classrooms] is making the greatest gain and which one isn’t, and then I try to analyze why that’s happening . . . it’s just something I monitor carefully.”

Of course, all three administrators do use the tests to monitor student progress, as well. The administrators spent a lot of time discussing how the test helps identify students’ “strengths and weaknesses.” As seen in Table 2.1, these strengths and weaknesses are sometimes in regards to individual student progress, but, at other times, they are viewed more holistically in terms of classroom achievement or instructional outcomes. The administrators explained that they use their understanding of these strengths and weaknesses to make decisions (generally) and to plan instruction (specifically).

Additionally, the administrators identify several other valuable uses for the MAP data (also seen in Table 2.1). All administrators mentioned the measurement and monitoring of gains over time – both in terms of academic gains during the school-year as well as gains and losses from one grade to the next. The administrators also report using MAP data to identify those students who will receive interventions, particularly those who need additional time and instruction to prepare for the state assessments. These students are often targeted through ability grouping during their regular instruction and may also be pulled out of other instructional time to be given additional preparation for the state tests.

Table 2.1

The Ways Administrators Report that MAP Data Reveal Strengths and Weaknesses

Strengths and Weaknesses	Administrator Quotations
<i>. . . of Individual Students</i>	<p>What are the students' strengths and weaknesses? Because some of them may be low, but they may be actually higher in some areas. Or they may be higher overall but they actually do still have some weaknesses in different areas. So, just looking at the different areas of the MAP really helps us.</p> <p>The teachers and the ICs [instructional coaches] look at individual student growth and areas of strength and weaknesses.</p> <p>The MAP testing gives us hot data, right now. What the children know and what they do not know.</p>
<i>. . . of the Students, Collectively</i>	<p>I have these [class] charts up here. We determine what areas are our weak areas [and] . . . where did we make our gains?</p> <p>It is a starting point for us to make decisions about what we should be teaching, where the students are.</p>
<i>. . . of Instruction</i>	<p>[The teachers] are able to discuss the areas, within their classroom, that seem to be strong, that seem not to be strong. So they're able to put their heads together and strengthen their instruction in the classroom.</p> <p>We try to look at what we seem to be teaching well and what we seem to not be teaching well and then I do staff development around it.</p> <p>What we want is to be able to analyze what students can do and cannot do, so that teachers use this information when they go back in the classroom to plan their instruction.</p>

The state assessments were a recurring theme in the administrator interviews (as they were with the teachers). In explaining the overall purpose for the MAP, one of the instructional coaches stated that the MAP is important because “we have the practice – in third, fourth, and fifth – we use this as a practice test for the state test.” That same administrator also explained

that the intent of the MAP for them (and the district) is really to predict the outcomes on the state test, the KCA: “My question has always been, ‘How reliable is it as a predictor of how they’ll do on the state tests?’ Because that is really its intent.” While this administrator seems less convinced in the reliability of the alignment between the MAP and the state tests (the KCAs), the other two administrators spoke of the importance of this alignment in outcomes, as seen in Table 2.2. This table also shows some of the other concerns about lack of alignment in the format of the questions, specifically the way material is presented and tested differently between the MAP and the KCA.

Table 2.2

Administrator Views of the Alignment (and Lack of Alignment) between MAP and KCA Tests

Alignment of Scores and Outcomes	Concerns about Lack of Alignment in Formatting and Questioning
“There is a correlation between the MAP test and the state assessment. It may not be obvious at first, but it seems like the kids that do a certain way on the MAP also do well on the state assessment. So there does seem to be a correlation between the two.”	“I still question what the reliability is to our state test. Even though I’ve been assured it’s high, I have my doubts sometimes . . . when I look at the state prep tests, their format is different.”
“It’s not an exact alignment with the Kansas state assessment . . . but we have to take the information from the MAP tests because that gets into the individual skills . . . And we have to focus in on . . . the individual skills so that we can get them prepared in those skills . . . and then have them prepared, as prepared as possible, to take the Kansas state assessment test.”	“I think I would want [the MAP] to align the question on the tests so that they are more in alignment with the questions on the Kansas state assessment.”

One instructional coach also introduced another factor that relates to the reliability of the mathematics MAP scores. As she described it:

The MAP math test is heavy on reading. And that is always an issue, but with the ESL kids, they struggle with it. And with kids who are low readers, they freeze because . . . the test is not read to them . . . And so, even if a kid could do the computation, he can’t figure out which numbers to put together to do the computation.

In other words, this administrator is questioning if a mathematics score is reflective of the mathematics abilities of the students if it is influenced by the language barrier. In some ways, this may help the reliability between the MAP and KCA since the KCA is also fairly text-heavy, so the MAP score alignment to the KCA results may not be affected by this concern. But the larger questions of what is being measured and if the assessment is providing accurate formative data for mathematics instruction remains.

Other concerns included the teachers' lack of familiarity with the tests. For example, one instructional coach pointed out that teachers "really can't go in [to see the test] unless they go in under a student and pretend they're taking the test." Similarly, the other instructional coach would like to see more specific data on the types of questions that students miss: "We know we need to focus our instruction around numbers and computation, but we don't know what strand, so . . . it's rather general."

Despite the variations between the tests and a few general concerns, the administrators were overwhelmingly in favor of continuing to use the MAP, defending its use during the optional Winter administration of the test, and identifying strong benefits to its use. There was a general consensus that more data are always good. For example, one interviewee offered, "[I]t does keep us focused. It's data. And so we can make decisions more objectively because the numbers are there." The principal was the most supportive of the continued use of the MAP, explaining: "Right now I think it's the closest thing that we have to getting us information about those students, about what we need to teach in the classrooms, about what we need to plan for in staff development for our teachers . . . It's a direct correlation between the planning, the instruction in the classroom, and how well the children do on the test." The other administrators concurred, stating, "I would continue using it just because it is really good showing progress"

and “I am happy with MAP.” To get a full picture, however, of how the MAP test is being used and received, one must turn to the perspective of the classroom teachers.

Qualitative Results: Why do teachers believe these tests are administered?

While the district and NWEA state that these tests are used to help classroom-level instruction, the teachers see the purpose of these tests to be more external to their work. The majority of the teachers talked about the desire of the district to have more data, in general, and to be able to measure growth. Three also talked about the district’s desire to have a “big picture” view of how its students compared nationally.

Two more interesting categories also emerged from the teacher interviews: the pressures of Adequate Yearly Progress (AYP) and the pressures of teacher accountability. Four of the teachers discussed the district’s desire to predict if the schools would meet AYP. One teacher commented, “I guess we use it as an indicator to tell if they’ll pass the test or not,” and another, more cynically explained, “I think it’s because if your school is not meeting AYP, they want to see it three times.” At least two of the teachers saw MAP testing as a means the district uses to hold teachers accountable. When asked about why she thought the district uses the MAP tests, one teacher said “[T]hey think it’s up to the teachers” and another noted, “[T]hey want to, of course, add in that whole teacher accountability . . . I really feel that some it’s just the pressure of performing for that.” While this aligns neither with the MAP’s official goals nor the stated goals of the district for MAP, it does reflect the same attitude revealed by the administrators at this school.

Qualitative Results: How do teachers perceive these tests?

Comparison of MAP to State Assessment. The teachers receive assessment data in many forms throughout the year and, therefore, as they discussed the MAP tests and resulting data,

they often did so in comparison to the other available assessments and data that they use.

Overall, the teachers concur that data from the MAP testing are not as directly applicable to instruction as the Kansas state assessment, known as the Kansas Computerized Assessment (KCA), and the practice KCAs. The teachers report much more familiarity with and applicability for the practice KCAs that this school uses weekly than they do with the MAP assessments. Therefore, several of them noted that the MAP serves primarily as “practice” with standardized testing and “exposure” to more types of questions and computer-based testing.

The teachers also realize that, while they are expected to use the results of the MAP testing, they are really only held accountable to the state tests and, in particular, to making adequate yearly progress (AYP). Several teachers made note that the MAP was much less important than KCA because it does not have an effect on whether the school meets AYP. One teacher put it bluntly, “AYP is not dependent upon MAP. I hate to say it but it’s not, so it’s not a focus.” Another teacher was even more adamant, “We get right into testing mode, and then it’s all about the state assessment because that determines AYP whereas MAP doesn’t. I [hope] that it doesn’t ever determine AYP, because we’ll be screwed having all those tests determine it.”

Comparatively Less Use of MAP Data. In contrast to the KCA data, which teachers report using to plan instruction, MAP data are not as easily applied to the classroom. Table 2.3 gives examples of teacher comments comparing the use of MAP testing results and KCA results. In these cases, the teachers are referring to “KCA” data in terms of the yearly state test results as well as the practice KCA test data that they receive throughout the school year. As seen in the table, MAP data are something the teachers mention “having,” but KCA data are those data that they more often report “using” in their planning and teaching. They explain that, in comparison to MAP data, KCA data are more directly applicable to their daily practices and to their general

understanding of how to help their students be “successful,” particularly when “success” is measured in terms of the yearly state test that determines AYP for the school.

Table 2.3

Comparison of MAP Tests (“Having Data”) to KCAs (“Using Data”)

<i>Having Data – MAP Tests</i>	<i>versus</i>	<i>Using Data – KCAs</i>
<p>“[MAP] doesn’t really show growth in the things that we test for in the state.”</p> <p>“The MAP testing, for me, is just awareness of your ability level . . . The MAP testing is just an identification . . . The MAP testing does not show you how to get better. It just shows you where you’re at.”</p> <p>“I don’t really use the MAP assessment to help me . . . I know they say the [MAP and state assessment] closely align but I don’t see it. I don’t see it in the scores, I don’t see it in the test that I’ve seen.</p> <p>“We don’t use the MAP data. We use KCA.”</p>		<p>“The KCA . . . now that has definitely changed the way I’m going to teach . . . next year.”</p> <p>“KCA – it’s more teaching . . . what students need in order to be successful.”</p> <p>“We look at where the kids are low, but, you know . . . we just tend to go towards whatever the KCA has given us and those tools, and focusing on that and not so much with the MAP.”</p> <p>“The [practice KCAs] are things that help you develop and improve . . . they show you where they’re at, but they’re also geared to show them how to get better.”</p> <p>“I look at their KCA scores, their practice tests. Because that is more closely related to the state assessment than MAP.”</p>

Beyond viewing MAP as just another experience with testing, however, most teachers associate MAP testing specifically with the data it provides them – data that some ignore, some value, and some fear. For example, some teachers do not find the data provided from these tests to be at all useful to their planning. As one teacher explained, “[W]e have data coming out of our ears and this is just one that kind of gets shoved to the side.” Others, however, did see some added value to these data. The majority of teachers discussed using the MAP results in some formative manner, including looking for student improvement or identifying students’ strengths and weaknesses based on the results of the MAP testing (or, often, based on the results of the MAP testing alongside the practice KCAs). About half of the teachers discussed general

changes in instructional material and/or strategies based on the results, but only two provided specific examples of such changes. A few teachers also mentioned using the results to identify what to “reteach” or how to group students.

Overall, however, the focus of many of the teachers’ comments was more on obtaining data rather than using them. One teacher explained, “The data for the MAP is confusing, ambiguous, and there’s a range of numbers, then there’s sets of different ranges, and then there’s a RIT score that has a range, and there’s another score . . . No, it does not help me at all. It gives administrators and me . . . some kind of vague category. So basically, we’re crunching data, we’re not looking at teaching strategies on that one.” In other words, the teachers know that they have these data and they do work with them in their collaboration times as is required of them, but these data are not really applied to their teaching (at least not to the same extent that data from the KCAs and practice KCAs inform their instruction).

General Concerns about the MAP Test. The teachers presented many other concerns about MAP testing, mostly around the vagueness of the data they receive. The teachers attribute this vagueness to their own lack of familiarity with the test but also to the structure of the data reported. For example, teachers explain, “I don’t know exactly what they’re testing” or “I don’t know what’s on the test. I don’t know.” But the teachers also correctly identify that they “don’t know . . . what section . . . that they [the students] are weak in. It doesn’t really say . . . This is all they give me for the MAP testing. It doesn’t break it down and explain.” Another teacher said the categories reported on the MAP are “too vague and not helpful” and added that “the test needs to be aligned with the curriculum, and the data in the tested indicators should be decoded, so that the common teacher can know exactly what indicators [are] being tested. And then the

math that involves all these various ranges, which are incredibly inconsistent – there’s a logic to it, but it’s unclear to me.”

Some teachers also lack trust in the validity and reliability of the tests – from students having bad days, to those who may just be guessing on the multiple choice items, to the influence of cultural bias and different forms of background knowledge, to ways that low reading skills may interfere with mathematics scores. One teacher stated, “I don’t think it really says what it’s supposed to say. But what can we do? We don’t have a choice.” And, while most teachers understand that the test is individualized to the level of each student, they are also cautious of this method of testing: “If they can answer hard questions, then the questions get harder. But if the questions are difficult for them, then they give them easier questions. I don’t know if that works or not. I’m not sure.”

Several teachers also object to the format of the test. For example, one teacher expressed concern about the length of the test and the shortness of the regular testing window (although the students can work beyond this window):

I think it’s wrong that they give them 57, 50 questions, or something like that and our kids don’t always get finished because we only get an hour in the computer lab to get it done, so some rush through because they’re tired. And they let them stay, but you know how some kids when they see everybody walk out the door, then they just go – they’re through, you know and then they just, you know, I think that sometimes has an effect.

Other teachers echoed these objections to the length of the test in terms of the number of questions the students must answer and the amount of time it takes to complete the test.

Additionally, some teachers did not like that the test consists exclusively of multiple choice items.

Inability to Teach to the Test. Many teachers saw the inability to teach to the test as a major flaw, along with the lack of alignment between the test and the curriculum. For example,

one teacher noted, “[T]he way that we teach math is not the way that it’s tested,” and another teacher commented that “we haven’t covered those things because it’s not a state assessed goal.” During the administration of the MAP, one fourth grade teacher, reflecting on the questions she saw her students trying to answer, commented, “He had spinners on there. We’ve never done spinners. I don’t even think third-grade does probability anymore.” The MAP test is designed to ask questions that are beyond (or even below) the grade-level standards for the students, however, because it is supposed to assess their current levels of attainment, generally; students who are proficient with grade-level material, therefore, should be receiving questions that are beyond the scope of their regular curriculum. This is precisely why one teacher feels frustrated with the MAP test, stating, “I think it’s harder for us to teach to the MAP test because it changes the questions according to how the kid is answering them. So how do we teach to the [standards]?” The teachers do not know what material to expect will be presented to each student on the tests both because of their lack of familiarity with the test and because of the test’s purposeful design to adjust to the level of each student.

An interesting dichotomy presented itself with these comments. Of the teachers who agreed that there was not a good alignment between the test and the curriculum and expressed desire to be able to better prepare their students for the MAP test, three also specifically mentioned that they value the authenticity of the MAP scores precisely because they feel that they cannot teach to this test. (Please refer to Table 2.4.) These teachers do not feel that they can (or do) teach to the MAP test as they do to the KCA and, therefore, they trust that the students performance on the MAP is more representative of true abilities and level of knowledge, not just good test-taking tips and tricks. These three teachers, therefore, actually named this feature of the MAP test as a benefit to their understanding their students’ strengths and

weaknesses. These teachers desire a new type of data that they can trust precisely because these data measure what they cannot directly prepare their students to mimic, but they also want to equip them to be successful and they do not know how to ensure that without some form of “teaching to the test.”

Table 2.4

Dichotomy: Desiring Authenticity but Pressured to Teach to the Test

	Valuing the Authenticity of Not Teaching to the MAP Test	. . . But Still Desiring to Teach to the MAP Test
Teacher A	“I think it . . . gives a clear picture, because I don’t teach to it. . . I think it’s also a truer picture of what they can do independently.”	“We don’t know how they’re going to ask the questions . . . so it might be good if we had question stems, so we could create our assessments . . . just for in-class . . . so that they’re familiar with [it].”
Teacher B	“I also like it because you can’t teach to the test . . . When you teach to the test you’re given what the question will look like and all the concepts, so you have a tendency to teach to the test because you’re worried about formatting. Well, with the MAP you don’t have that. At least I don’t, because I’ve never looked at it . . . And so . . . it’s really what I’ve taught . . . It makes me teach children to think more and to reason because anything is expected.”	“Well, for me, it actually does help keep me on track because everything on the KCA is not all of our benchmarks that we’re supposed to meet. So it’s like the MAP helps me keep on track . . . if you keep your focus on testing.”
Teacher C	“We’re just encouraged to teach to the tests, which is not good. And if we don’t know what’s on the MAP test, it’s hard for us to teach to the test.”	“We’re just not sure how to go about preparing the kids for the MAP test . . . They’re putting us under pressure to teach to the test . . . So if I could kind of see what the test items were . . . that would be nice.”

For these teachers, the MAP score is a more valid number than a KCA score, but that does not mean that it is a more useful number. Teachers still report much greater use of the KCA and practice KCA data because they directly relate to the state standards. Teachers have a better grasp, then, on how the KCA data should inform their instruction. They more clearly see how to

use the practice KCA results to help boost their students' performance on the official state assessment (the KCA), and this is the number from which AYP is determined. While they have to be most concerned with KCA scores because of the policy pressure and implications that result from AYP, as educators who are interested in the ability-level of their students, they trust the MAP scores' representation of ability levels much more than they do the KCA scores.

Overall, the teachers prioritize the data they receive into what is most useful for them for their instruction. The teachers unanimously report using KCA data to make decisions on instruction and most report the ease with which the KCA tests match the state standards and, to a certain extent, their curriculum. Even though none of the interview questions asked about the KCA, the teachers intuitively compared MAP testing and data to the KCA's testing and data to explain the difficulty that they have in applying results from the MAP testing to their classroom. The teachers desire to have more data; they buy into the value of data. But they find the MAP data vague and confusing. They do not want to have to dig or sort through the data in order to find something useful. If data are not clearly showing them what to do in their classroom, they put them aside. Even the few teachers who expressed a positive view of the MAP data over the KCA (because of its "authenticity") still want to find a way to prepare their students more for the test, i.e. to teach to the test. They want their students (and therefore, themselves) to be successful; and success has been defined to them repeatedly as good test scores.

Quantitative Results

The descriptive statistics from the teacher surveys also demonstrate how the teachers perceive the MAP assessment. As seen in Table 2.5, almost all of the teachers agree (or even strongly agree) that the tests are viewed as a reflection on their teaching. So, while only two

teachers mentioned this as a purpose of the district in their interviews, all but one of the teachers recognized this as a pressure when they were asked about it directly in the surveys.

Table 2.5

Results from the Teacher Survey regarding the NWEA-MAP Assessment

Survey Statement about the NWEA-MAP Assessment	Percent Agreeing
Administrators in my school believe students' NWEA-MAP assessment scores reflect the quality of teachers' instruction.	91**
The NWEA-MAP assessment has brought much needed attention to education issues in my school.	73
Overall, the benefits of the NWEA-MAP assessment are worth the time invested.	73
The NWEA-MAP assessment is <i>as accurate</i> a measure of student achievement as a teacher's judgment.	64
NWEA-MAP assessment is NOT an accurate measure of what my students know and can do.	55
Taking the NWEA-MAP assessment is a good use of student time.	55
Scores on the NWEA-MAP assessment accurately reflect the <i>quality of education</i> students have received.	46
The NWEA-MAP assessment is just another fad.	46
The NWEA-MAP assessment is compatible with my district's mathematics curriculum.	46
The instructional texts and materials that the district requires me to use are compatible with the NWEA-MAP assessment.	36

**Two strongly agree.

Overall, the teachers reject the idea that these tests reflect the quality of education that the students receive. And, as was true in the interviews, the teachers are concerned that the curriculum and, in particular, the texts and materials the district prescribes are not compatible with the MAP test.

The teachers do, however, see value in the assessment. Nearly three-quarters (73%) agree that this test brings attention to issues at their school that need to be addressed, and the same percent also agree that the benefits are worth the investment, although only 55% believe it is a good use of the students' time. The teachers were less certain about, but still favorable

toward, the accuracy of the data they receive from MAP. Sixty-four percent believe that “the NWEA-MAP assessment is as accurate a measure of student achievement as a teacher’s judgment.” Yet the majority also agreed that the test “is *not* an accurate measure of what my students know and can do.” Similar to the teacher interviews where teachers value the authenticity of scores yet want to teach to the test to increase scores, this contradiction in survey responses may indicate another dichotomy in the understanding of the MAP testing. Yet, it is possible that this is not a contradiction at all but rather an indicator of a lack of confidence in their judgment as teachers. If the teachers do not think they have control over the outcomes of their students’ achievement levels, then they could agree with the first statement and with the second without contradicting themselves.

To explore this possibility, an examination of the teachers’ levels of mathematics teaching efficacy and outcome expectancies as measured by the MTEBI is helpful (Enochs, Smith, & Huinker, 2000). The twenty-one MTEBI items included on the teacher survey allow analysis on two sub-scales: outcomes expectancy beliefs and mathematics teaching efficacy beliefs. Outcomes expectancy beliefs, those beliefs in the likelihood of the consequences of one’s efficacious actions actually occurring (Bandura, 1986), may be most applicable to these teachers’ views of the MAP tests. As Tschannen-Moran, Woolfolk Hoy, and Hoy (1998) explain, “The efficacy question is, Do I have the ability to organize and execute the actions necessary to accomplish a specific task at a desired level? The outcome question is, If I accomplish the task at that level, what are the likely consequences?” (p. 210). In this case, the MAP scores are the outcomes or consequences.

The average teacher efficacy and outcome expectancy levels are shown in Table 2.6. The outcome expectancies are statistically significantly lower than the teaching efficacies ($p < .01$).

And these low outcome expectancies help explain the apparent contradiction in the teachers' view of the MAP result.

Table 2.6

Teacher Efficacy and Outcome Expectancy (n=11, scale 1-4)

MTEBI Subscale	Mean	Minimum	Maximum	Standard Deviation
Mathematics Teaching Efficacy	3.31	2.85	3.92	0.38
Outcome Expectancy	2.81	2.25	3.25	0.31

The teachers see the measurement of student achievement – based on their own judgment and by the test – as something over which they have less control, even when they believe that they have the ability to teach mathematics well. The teachers can, therefore, hold the opinion (as the survey reveals that they do) that the test is not an accurate measure of what their students know if they believe their students actually know more than the test reveals. And teachers can simultaneously believe that the test measures their students' achievement as accurately as they, the teachers, can judge the students' achievement because they feel less control over the outcomes (i.e. the student achievement levels) than they do over their ability to teach mathematics well.

Analysis

The official view of the district is that the MAP assessments provide information on student strengths and weaknesses by identifying their current instructional levels. This information is primarily intended to be formative in nature for classroom teachers to monitor progress of individual students and their classes as a whole. As explained by a district administrator, the MAP test should provide useful information on “mapping” how to get the students from their current instructional levels to where they need to be. All groups – NWEA,

this school district, the school-level administrators, and the teachers – agree that the MAP test is intended to provide data that can be used to inform instruction. However, when one of the district administrators refers to getting to where “the student needs to be,” one can assume, based on the administrator interviews, that his comment can really be interpreted to mean “passing the Kansas state assessment.”

The school-level administrators, in fact, comment directly on this goal of the MAP testing. They see a great value in the MAP data’s ability to predict KCA scores, although they question how well-aligned MAP scores are to the KCA scores. NWEA has produced reports on the alignment of these scores for most states, including an updated report in November of 2009 for the alignment of MAP RIT scores with Kansas state assessment scores (Northwest Evaluation Association, 2009c). In these reports, NWEA provides charts with minimum cut scores by grade-level, i.e. RIT scores that a student must reach in order to score within a given achievement level on the state tests (such as “academic warning,” “approaches standards,” “meets standards,” “exceeds standards,” or “exemplary”). These charts can be used to identify individual students who are “at risk” for not passing the state assessment. Therefore, NWEA is aware of and aids states and districts in using the MAP testing for the purpose of score prediction for the state assessments.

The consensus on these purposes of MAP testing, however, is not surprising; these are the openly discussed purposes of which everyone in the district involved with the MAP should be aware. What is more interesting is the underlying, often unstated goals of MAP testing as well as how these purposes are interpreted and potentially fulfilled at the level of implementation – the classrooms. The teacher and administrator interviews provide insight into these unofficial goals.

Unstated Purpose: Teacher Evaluation and Accountability

The most remarkable result of analyzing the district purposes of the MAP tests came from the school-level administrators' interviews. These interviews revealed how the administration makes evaluative judgments of teachers based on MAP results. Surprisingly, the administrators were very open and transparent about discussing the use of the MAP in terms of how they judge teachers and keep them accountable. While this is not an official purpose of MAP testing – based on district and NWEA documents – it is a well-accepted purpose at the level of implementation, in the schools.

The teachers also discussed the evaluative uses of the MAP, indicating that they are aware of the ways the results reflect on their teaching. There is an important difference, however, between the response of the administrators and that of the teachers to the use of the MAP data for teacher evaluation and accountability. The administrators see the evaluative use of the MAP data as helpful in that it can motivate teachers to do better and in that teachers can also be given extra support and professional development if their weaknesses as instructors are identified in this way. The teachers, on the other hand, see this evaluative purpose as punitive. As will be explored in Chapter Three, this awareness of being judged by their administration based on the MAP results leads to increased stress for teachers and an internalization of the pressures of testing.

Mistrust of MAP Results

As seen in this chapter, the teachers' frustrations also stem from their mistrust of the MAP results. The teachers are concerned that an increase in the scores on these tests may only demonstrate better test-taking skills, not better mathematics skills. However, some teachers believed that MAP testing had an advantage over the state testing because they cannot teach to

the MAP assessments as well. Therefore, the teachers saw MAP testing as more conducive to revealing accurate student achievement (compared to the KCAs) because they feel that they can teach directly to the content and form of questions on the KCAs but do not feel familiar enough with the structure of MAP to be able to teach to the test. These teachers want to maintain idealism where education and learning are the priorities over test scores, and yet they are forced to realize that they and their students are being judged by these scores and, therefore, they do not want to be hurt by them.

One of the administrators also emphasized a mistrust of MAP results specifically for the large population of English language learners at this school. Since the MAP tests are “wordy” and text-heavy, and since, after second-grade, the test no longer includes the audible reading of questions to the students, the students’ mathematics scores may reflect their language ability more than their mathematics ability. Despite this realization by the school administrator, all of the teachers have been instructed that they cannot read aloud any word or question to the students while students are taking the MAP test.

For these reasons, it is unsurprising that the surveys revealed that the teachers do not feel that they have much control or influence over the student outcomes on standardized tests and, specially, on the MAP results. This is not based on their lack of confidence in teaching mathematics well (although some of them do express concerns in this area, at times). Instead, their survey results revealed that their mathematics teaching efficacy is significantly higher than their outcome expectancy, indicating that even when they feel confident in their ability to teach mathematics well, they do not feel that this translates directly into good test scores for their students. And, since the teachers feel that their ability to teach is being judged by these scores, their level of frustration toward and mistrust of these tests is even more understandable.

The Assumed Value of Obtaining More Data

Despite some discrepancy in the credibility of MAP results, there is consensus among the administration and the teachers that MAP testing is providing more data. The system as a whole and the individuals within the system (administrators, teachers, and even the students) are operating under the assumption that more data can only be good. The teachers and administrators want to make informed decisions and want the data in order to have greater confidence as they make decisions. These decisions take place at the level of the individual (pulling students out for intervention, assigning students to small groups), the classroom (identifying areas for remediation in instruction, informing instructional content and practices), and the school (school-wide test preparation strategies, comparing classroom outcomes, evaluating teachers). As a general rule, the teachers and administrators believe that by increasing the amount of data available to them, they also increase their effectiveness as educators. Yet, as seen in the interview results discussed in this chapter, the focus is greater on obtaining and having the data than it is on interpreting and using the data.

Therefore, several questions must be raised to challenge the assumption that obtaining more data is actually benefitting the students and the school. The subsequent chapters address some of these key questions. For example, Chapter Three will explore the intended and unintended effects of obtaining and using these data, and Chapter Four presents evidence about whether MAP data present new information about student achievement that the teachers and schools would not otherwise have. The data in this chapter help in looking at another key question about the value of more data by providing insight into whether the data are being used as they were intended. In particular, the results in this chapter allow for an examination of the

utility, or usefulness, of the MAP data at the school-level, specifically its utility in informing instruction and providing guidance for choosing intervention strategies for students.

The Utility of MAP Data Compared to KCA Data

Despite the consensus that teachers and administrators are happy to have additional data supplied by MAP testing, the data themselves are a source of great frustration for the teachers since the data, particularly the RIT scores, are difficult to interpret and even more difficult to apply directly to their instruction. In fact, the teachers and administrators commented on the ease of using the practice KCA results, rather than MAP, to identify specific areas for remediation in instruction. The teachers, in particular, repeatedly compared the utility of the data provided by MAP to that of the practice KCA data.

The comparison of the MAP to practice KCA data is most illuminative of the teachers overall frustrations with MAP. The teachers realize that they have an abundant amount of data on student weaknesses and strengths, especially since they receive weekly formative data from the practice KCA tests. While nothing in the interview questions asked about the practice KCA data, the KCAs were mentioned repeatedly when the teachers were asked about how they use the MAP data. They wanted to discuss how data inform their instruction, but their examples of regularly using data in this way come only from the practice KCAs, not the MAP. They expressed their frustration with the density and complexity of the MAP reports, the confusion over the meaning of RIT scores compared to percentile scores, the vagueness of the reported sub-categories of mathematics, the lack of alignment with the curriculum, and their lack of familiarity with the MAP test questions and format. Therefore, when asked about why they believe the MAP test is administered, the teachers did not speak about the need for more

formative data for planning their instruction. Rather, they often pointed to the desire of administrators to predict the results of the state test and, from this, predict AYP.

Notably, the charts provided by NWEA for predicting KCA results from MAP results explain that “meeting the minimum MAP cut score corresponds to a 50 percent probability of achieving that performance level” (Northwest Evaluation Association, 2009c, p. 5). So, when administrators and teachers use these cut scores, they are actually presenting a score that gives only a fifty-percent chance of a student then passing the state exam. This predictive value seems low, and is probably much lower than the teachers and administrators would assume it is when they use it to predict AYP.

But an even greater concern is that, when used in this way, the MAP test can be considered formative only in that it can become a red-flag of possible failure to meet AYP and point to a need for intervention. The teachers were clear in explaining that these tests do little to pinpoint *what* intervention is necessary, relying instead on the practice KCAs to identify specific standards and areas for remediation of instruction. So, from the perspective of the teachers, the MAP test is actually being used as an interim summative assessment to predict what the high-stakes summative assessment will show.

In terms of practically informing instruction, MAP may only serve to confirm the data that teachers and administrators receive from the practice KCAs; this will be further explored in Chapter Four. Before this, however, the potential drawbacks and consequences (intended and unintended) from the implementation of MAP testing and the use of MAP results will be examined in Chapter Three. If obtaining “more data” is in fact a key “good,” or benefit, that teachers and school administrators derive from MAP, the other effects of MAP testing must be weighed against this “good” of more data.

Chapter Three

What are the Effects of This Additional Testing?

While the MAP testing does provide the district and administrators with an indication of student progress and predictions for state assessment scores, the stated value of the assessment is its ability to measure student growth over time and to provide formative data for classroom teachers to use in planning instruction for their students. This chapter will explore the effects of the MAP testing at the classroom level. Specifically, the data in this chapter will help to explain how additional mathematics testing and data obtained through the MAP influence instruction in terms of feedback and evaluation of students, instructional planning and material selection, as well as intervention with students and student grouping.

Yet, instruction is not a factor that should be viewed in isolation. Instruction is an activity that occurs in classrooms between two parties – teachers and students. The classroom and school environment created around testing affects people first, and through its effect on these individuals, it secondarily affects instruction. Therefore, the most important question on the effects of MAP testing is how this testing affects the teachers and the students. This question is at the heart of the data and results presented in this chapter.

Data

To explore the effects of MAP testing on teachers, students, and instruction generally, multiple forms of data were collected. Separately, these data provide little concrete information on the effects of MAP, but they piece together a picture of the instructional environment, the individuals teaching and learning within this environment, and the roles that MAP testing plays in shaping the environment. Therefore, this chapter includes multiple sources of quantitative and qualitative data.

The quantitative data come from the teacher surveys. Ten questions from the survey on the ways that teachers use the MAP data are presented. Teachers were also asked about the amount of pressure they feel from various sources related to the MAP results. These sources and the perceived amount of pressure from each source are included in this chapter.

The qualitative data, however, are the fundamental components of this chapter. Classroom observations and, in particular, the observations of the administration of the Spring Assessment of the MAP are explored. Since every third, fourth, and fifth-grade class was observed prior to testing, during testing, and after testing, the observation notes are key in presenting the environment surrounding MAP testing at this school. Additionally, comments from the administrator interviews and the teacher interviews offer further insights into the effects of MAP testing on the teachers, students, and instruction.

Methods

Survey Data

The data analyzed in this chapter include seventeen items from the teacher survey. Ten of these items related to the ways that the teachers report using the MAP data, and two items examined teachers opinions of student reactions to MAP testing. These twelve questions were adapted from a national survey used in a study by Pedulla, et al. (2003). Please refer to the Methods section in Chapter Two for a further description of how items were selected and adapted from the survey of Pedulla, et al. for use in this study. The results of these four-point Likert-scale items are analyzed in this chapter as descriptive statistics, with the percent of teachers agreeing (or strongly agreeing) presented in the tables.

The other five survey items included in this chapter asked teachers to indicate the amount of pressure that they perceived about MAP results coming from five different sources. These

questions were adapted from a survey item used by Moore and Waltman (2007) in their study about pressures related to increasing test scores in response to NCLB. Because these items were not presented with a Likert-scale, they were included in a separate section of the survey from the other items analyzed in this chapter. For these five items, teachers were asked to indicate the extent to which they feel pressured by each source to increase students' scores on the NWEA-MAP assessment. Teachers then circled "a lot," "a little," or "none" for each of the five sources listed. These results are presented in terms of a percentage of teachers indicating each level of pressure from each source.

Observational Data

Classroom Observations Prior to MAP Testing. Every classroom teacher was observed teaching a mathematics lesson prior to the MAP testing. These observations occurred in the few weeks between the completion of the state KCAs and the beginning of the MAP testing. The teachers were aware that I would be coming into observe at some point during those two weeks. They provided me with their regular daily schedule so that I would know when they would be teaching mathematics. They also let me know ahead of time if there were any field trips or other days with changes to their regular schedules so that I would avoid those days for observation. Other than that, the teachers did not know when, specifically, I would observe. However, I always asked permission of the teachers when I entered to room to be sure it was a "good time" for them. The teachers were very receptive to my visits and would find a space for me in the back of the room when I came in. They often apologized about starting the lesson a few minutes late or doing a review activity. I assured them that whatever they were doing that day was fine; I was just there to observe their normal classroom activities during "math time." In one case, a

teacher was preparing to have her room painted and was switching classrooms unexpectedly, so I had to return the next day.

Detailed observation notes were recorded in a notebook; these notes included quotations from the teachers and students, examples of problems written on the board, details of student-student and student-teacher interactions, types of activities, levels of student engagement, and other related comments. Additionally, formal observations were made using the CETP – Core Evaluation Classroom Observation Protocol (Lawrenz, Huffman, & Appeldoorn, 2002); notations on this protocol include recording the type of instruction, level of student engagement, and level of cognitive activity in five minute blocks. Space is also provided to summarize the lesson and its purpose and to evaluate overall indicators of lesson effectiveness. (Please refer to Appendix D for this pre-MAP testing classroom observation form.)

Observation notes were read and re-read by grade-level to look for commonalities and differences in the lessons. For example, the commonality of vocabulary development was found within and between grade levels. References to standardized testing during the lessons were also found in many of the lessons. Examples of these commonalities and differences were highlighted and compared with each other; related quotations and antidotes, especially those that seemed to best illumine data from the surveys and interviews, were then included in the data presented in this chapter.

Observations of MAP testing. While I was able to be present in the classrooms for entire mathematics lessons, I was not always able to remain in the computer lab for the entire time that a class was testing because several grade levels were often testing at the same time in different labs. Therefore, I made it a priority to observe the beginning of the testing for each class as well as the times when students were finishing. I also was careful to make sure that I spent an

extended, uninterrupted period of time (usually at least thirty minutes) observing each class while testing in order to see the events of the testing progress for each class. In a few cases, testing windows did not overlap and, therefore, I was able to remain with the class for its entire testing time. As with the classroom visits, observational notes were collected during MAP testing in a notebook. While no formal form was used for recording blocks of time, these notes included references to the time throughout in order to provide a structure to the notes and to give indication for how quickly or slowly some students would finish or move through questions.

In a similar manner to the analysis of the classroom observation data, the MAP testing observation notes were organized by grade level and repeatedly read and reviewed for themes, with commonalities or exceptional examples being highlighted or coded. Comments by teachers or administrators were also highlighted in a different color from student comments in order to better identify commonalities among teacher reactions and commonalities among student reactions since these data provide insight into both perspectives. For example, the commonality among the comments teachers made to their classes as they entered the computer lab to begin testing was only seen once teacher quotations were highlighted in the observation notes and then the notes were re-read. The results of that analysis are presented in Table 3.6. Observation notes from the MAP testing were analyzed separately from the classroom observation data, although, in the analysis, these data were often paired together in order to understand the fuller picture of the testing environment at the school.

Classroom Observations After MAP Testing. The classroom observations after MAP testing were done in a different format than those conducted prior to MAP testing. Rather than observation of full lessons, these observations were shorter (about ten minute) walk-through observations. The information provided in these shorter observations still gave a full picture of

whether lessons were continuing “as usual” or if there were differences in the styles of the lessons after the testing was complete. Since this was the main purpose in doing the post-MAP testing observations, these shorter walkthroughs were all that were necessary. Additionally, this was a practical decision. I did not want to observe classes during the last week of school since most teachers always change the format of their days and lessons at that point in the school year. Additionally, there were many other disruptions (such as field trips and assemblies) that had to be accommodated in the schedule of observing classes. Therefore, in order to have the opportunity to visit all the classrooms, multiple classes had to be visited during the same timeslots. The walkthrough observations, therefore, allowed for this flexibility to visit every class.

Notes on these walkthrough observations were recorded on an observation form that included information about levels of student engagement, primary teaching methods, level of cognitive activities, clarity of lesson objectives, and other comments on the lesson. This form was based on one developed by a peer teaching observation program (Parker High School, 2009). It is also consistent with the recommend use of walkthrough observations (Ginsberg & Murphy, 2002; Richardson, 2001; Skretta & Fisher, 2002). (Please refer to Appendix E for this post-MAP testing classroom observation form.)

In analyzing the post-MAP assessment classroom observations, levels of student engagement and levels of cognitive activity were compared to those observed prior to MAP testing for each teacher. The post-MAP lessons were also categorized based on the types of material taught and level of cognitive activities included in the lessons. Based on the observational data, these categories included those not teaching mathematics, those reviewing grade-level material with low cognitive level activities, those reviewing below grade-level

concepts, and those introducing new material using low cognitive level activities. All of the observed lessons fit into these categories, so there was no need of additional categories such as “those reviewing grade-level material with high cognitive level activities.”

Interviews

Data from the teacher and the administrator interviews are included in this chapter. Please refer to the Data section of Chapter Two for a detailed description of how the interview data were collected and analyzed. Included in this chapter from the teacher interviews are categories and themes that emerged related to external (Table 3.4) and internal (Table 3.5) pressures on teachers based on MAP testing and results. Administrator perceptions of teachers’ reactions to MAP testing are then presented in Table 3.7.

Results

How does this additional testing and data affect instruction?

Qualitative Results. As one would expect from previous research on standardized testing (Abrams, Pedulla, Madaus, 2003; Pedulla et al., 2003; Smith, 1991), one effect of additional testing is the loss of instructional time. The instructional coaches seemed to be the most concerned with this effect. One stated, “[I]t does take time . . . We’re probably setting aside about three weeks now, just for MAP testing every single child.” The other explained, “[P]robably the primary cost is just instructional time . . . there are competing priorities for teaching time and it’s hard for teachers to decide.” She also talked about the time needed to go through the data after the testing: “I don’t think that we have had the time – we’ve not taken the time to really use MAP as thoroughly as we could.” So, the time factor is not just an instructional issue, as expressed by most of the previous research, but there is a struggle to find time for the analysis of data, along with the time for planning and preparation based on these

data. Additionally, MAP testing is seen as just one part of the time given to assessment. In response to a question about how much testing is done overall at the school, the principal simply said “a lot.” The instructional coaches said “[O]h my goodness . . . probably once a week [they] did some kind of a KCA test . . . so, wow, once a week” and “[A]t [grades] three, four, five, between the state test, the KCAs . . . and MAP and their testing it’s a lot . . . we need to be cautious about over testing the kids because they burn out and we’re not getting valid results.”

The classroom observations provide the strong evidence of the impact MAP testing has on instruction. Most of the teachers’ lessons leading up to MAP testing included an emphasis on mathematics vocabulary development. Teachers were also likely to make references to testing in their lessons. One fourth-grade teacher, who was doing a lesson on three-dimensional solids, told her students, “I’ll probably look . . . after you take the test to see how you do on your shapes.” A fifth-grade teacher was reviewing fractions and said, “We use the stuff that we learned for the state assessments to reduce fractions.” A student then asked, “Are we taking another state assessment?” The student (and teacher) seemed relieved that the state assessment was done for another year, but the focus was still on reviewing and improving these same skills as they prepared for the MAP. Most lessons involved higher-order thinking skills (such as application and synthesis) for at least part of the lessons.

The instructional environment changed dramatically after MAP testing was complete, with differences seen in the concepts being taught and the cognitive activity levels expected of the students. A summary of these differences is presented in Table 3.1. Also shown in this table is a general category of the type of lesson material being taught along with the levels of student engagement in the lessons, which, unlike the levels of cognitive activity, remain the same or increase after the completion of MAP testing.

Table 3.1

Summary of Pre and Post MAP Testing Classroom Observations, Including Topics of Lessons, Average Cognitive Activity Levels (C.A.), & Average Levels of Student Engagement (L.S.E.)

Gr.	Pre-MAP Testing Observation				Post-MAP Testing Observation			
	Topic	C.A.	Category of Lesson	L.S.E.	Topic	C.A.	Category of Lesson	L.S.E.
3 rd	Equivalent Fractions	1	“Scattered” with grade-level tested material	Medium	Three Dimens. Figures	2	New (non-tested) material	Medium
	Multipl. & Division w/ Cubes	2	Grade-level, tested material	High	n/a	0	<i>Math Not Being Taught</i>	n/a
	Multipl. Fact Families	3	Grade-level, tested material	High	Multipl. Relay Game	1	Low-level review of grade-level material	High
	Represent Fractions	3	Grade-level, tested material	High	n/a	0	<i>Math Not Being Taught</i>	n/a
4 th	Division Vocabulary and Word Problems	2	Grade-level, tested material	Medium	Division Vocabulary	2	Low-level review of grade-level material	High
	Three Dimens. Figures and Landscapes	3	Grade-level, tested & non-tested material	Medium	n/a	0	<i>Math Not Being Taught</i>	n/a
	Multipl. & Div. Activity (Plan a Picnic)	3	Grade-level, tested material	Low	Addition with Money Word Problems	1	Below grade-level material	Medium
5 th	Fractions, Percents, and Decimals	2	Grade-level, tested material	Medium	Perimeter Word Problems	2	Below grade-level material	Medium
	Ordering Fractions and Percents	2	Grade-level, tested material	Medium	Fractions and Percents	2	Low-level review of grade-level material	High
	Fractions Webquest	3	Grade-level, tested material	Medium	Ordering Fractions	2	Low-level review of grade-level material	Medium
	Simplify Fractions	2	Grade-level, tested material	Medium	Converting Fractions to Decimals	1	Low-level review of grade-level material	Medium

1=Receipt of Knowledge; 2=Application of Procedural Knowledge; 3=Knowledge Representation;
4= Knowledge Construction; 0=Math Not Being Taught

As seen in Table 3.1, all but one of the teachers were using higher-order cognitive skills in their lessons prior to MAP testing, and all of the topics of the lessons were grade-level appropriate concepts. (As noted in the table, one third-grade teacher presented a lesson that was “scattered,” often going off-topic on tangents unrelated to the lesson, although the main topic was grade-level appropriate material.) After MAP testing, however, only one teacher introduced new material to the students and some teachers were no longer consistently teaching mathematics. Specifically, during the walk-through observations after MAP testing, three teachers were not teaching mathematics at all, five teachers were doing very low-level reviews of old material from the year, and two teachers were reviewing below-grade-level concepts. Only one teacher was introducing new, non-tested material to her students, although even this was done using low-level cognitive activities.

Also, as seen in the pre and post MAP testing observations in Table 3.1, the levels of cognitive activity generally dropped in the lessons presented after MAP testing. For the three teachers whose lessons had the same average level of cognitive activity both before and after MAP testing, the level of the material presented in the lesson was still less challenging. The only exception, where the level of cognitive activity of the lesson increased between the observations, was seen in the third-grade teacher whose initial lesson was “scattered.” In the case of this teacher, it seemed that the lessening of pressure after the testing may have allowed her more flexibility to teach. Interestingly, the levels of student engagement remained fairly constant between the observations. And, in the three cases where a change was seen in the levels of student engagement, the levels actually increased.

One fourth-grade classroom can serve as an example of these trends in instructional changes. During the mathematics lesson observation prior to MAP testing, the class was

working on a multiplication and division activity where the students were asked to plan a picnic for twenty-four people. In this real-world scenario, they had to identify how many of each item to buy (where many of the items came with multiple items in a pack, e.g. eight hot dogs in a pack or twelve buns in a pack), find the total cost, and calculate the cost per person. This was a complex problem, in that there could be multiple “correct” solutions and the students had to manipulate the costs of the items in various ways to plan their picnic budget. The level of cognitive activity for this lesson was high (including both knowledge representation and knowledge construction). Unfortunately, the student engagement level was low. This seemed to be related to classroom management, however, due to multiple factors. Some of these factors were external to the teacher’s control, including that school-wide snacks were delivered during this time and that the mathematics lesson timeslot was cut short for a special “advocacy lesson” that the school counselor was going to present. Other issues that distracted students from focusing on the task were more specific to the classroom and this lesson’s structure; for example, the students were allowed to randomly choose a partner or to work on their own, and the timing of the lesson was rushed, with the teacher having them start this activity during this day but planning to finish the activity the next day.

On my second lesson observation in this classroom, after MAP testing, the students were working on short, simple word problems involving addition and subtraction with money. While the content, therefore, bore some similarity to the previous lesson, this lesson’s material was below-grade-level, involving only addition and subtraction in simple one and two-step word problems that had a single solution to each problem. The worksheet actually contained a few more challenging problems that involved the concept of equivalent fractions, but the teacher told the students that they did not have to do that part. Despite the lower cognitive level of this

activity (knowledge retrieval and some comprehension) and the material being below grade-level, student engagement did increase with noticeably more students on-task during this observation time. In part, this seemed to be due to fewer interruptions in the time designated for mathematics instruction that day. As reported in Table 3.1, these types of differences pre- and post-MAP testing were consistently seen in almost all of the third, fourth, and fifth-grade classrooms at the school.

Quantitative Results. Teachers report that the primary use of MAP data is the evaluation of student progress. As seen in Table 3.2, eighty-two percent of the teachers at the school use MAP for this purpose.

Table 3.2
Teacher Survey Results on Uses of NWEA-MAP Data

How Teachers Use NWEA-MAP Data	Percent Reporting Using Data in this Manner
Evaluate student progress	82
Assess my teaching effectiveness	73
Give feedback to students	73
Group students within my class	73
Plan my instruction	73
Give feedback to parents	64
Plan for remediation for students	64
Select instructional materials	64
Plan curriculum	45
Determine student grades	18

However, evaluating student progress does not mean a formal evaluation in terms of grading, as only two of the eleven teachers report using MAP scores to determine student grades. Instead, the teachers evaluate student progress by looking for growth from one administration of the test to the next (or, even more long-term, from the beginning of the year to the end of the year). The teachers also use the MAP data to assess their own teaching effectiveness, with seventy-three percent reporting the use of the MAP results in this way. Additionally, the majority of the teachers (73%) report that they use MAP results to give feedback to their students, i.e. to show students their own progress over time, and many of the teachers (64%) see MAP results as a source of feedback for parents, as well.

Beyond evaluation and feedback, seventy-three percent use the results of MAP testing for planning their instruction and for grouping students within the class. Similarly, sixty-four percent of the teachers report using these data to plan remediation for students as well as to select their instructional materials. Teachers did not see MAP data as readily useful, however, for planning curriculum, with only forty-five percent reporting that they use MAP data in this way.

What are the effects of MAP testing on the teachers?

Quantitative Results. Teachers feel a lot of pressure regarding standardized testing results. Certainly, this pressure is most acutely felt in terms of meeting AYP for NCLB. Since the MAP testing does not affect AYP status and since its main purpose is supposed to be to help teachers inform their own instruction, one might expect that teachers would not feel much pressure from this test. The survey results in Table 3.3, therefore, seem surprising. Only one teacher indicated not feeling any pressure in regards to the MAP testing. (Please refer to Table 3.3.) Interestingly, that teacher was also the only first-year teacher in the group and, therefore, she is possibly not yet aware of the weight that may be placed on these tests.

Table 3.3

Results from Teacher Survey on Sources of Pressure Related to the MAP Results

Teachers feel Pressure from . . . ?	Percent Indicating Each Level of Pressure		
	A Lot	A Little	No
Themselves	73	18	9
School District	64	27	9
School-Level Administrators	55	36	9
Colleagues	36	55	9
Parents	9	27	64

Of the categories listed in the table, the only place where teachers seem to get relatively little pressure toward MAP results is from the parents. The greatest pressure is from themselves, followed closely by the school district, then school-level administrators, and, finally, colleagues.

Qualitative Results. As was seen from the survey results (in Table 3.3), the teachers feel a lot of pressure related to the MAP testing. Some of this pressure is internalized; they see the students' scores as a personal reflection on their own teaching. However, the teachers also mentioned external pressures in their interviews, as seen in Table 3.4.

Table 3.4

Comments by Teachers on the External Pressures Related to Accountability and MAP Results

Really, they're grading the teachers.
I am the leader of the class and . . . [testing] could have adverse effects on me . . . and possibly undermine some of the things that I am trying to do.
It's [an] . . . assessment of how the kids have learned during the year <i>and</i> if the teachers were accountable for their learning.
It's all about numbers. It's not about learning.
Oh a lot of stress . . . And the teachers are put under pressure.
How does administration look at those scores in relationship to your classroom practice? . . . Last year, it was kind of a big deal . . . By the time we really sat down and began to look at our KCA and MAP RIT scores at the end of the year . . . you were really out there. And if it didn't appear that your class had made a lot of movement, it didn't look very good. I didn't like the way it was done.
What's hard is [that] you're disappointed [with the scores] and all of the sudden you've got to walk in and you've got to talk about it.
My class is the lowest. That's what I get out of [our meetings].

Much of this external pressure is framed in terms of being held accountable by the district through the test results. As shown in Table 3.4, over half of the teachers mentioned this form of external accountability in their interviews.

Some teachers also feel that the test itself is “unfair” and, therefore, that they are being unfairly penalized by the test. These teachers see standardized tests, in general, as biased and, therefore, believe that the MAP test is a biased test. They also see how it unfairly penalizes their school compared to schools with more homogenous populations and with less student turnover. One teacher explained that “it penalizes urban schools and special education students and English language learners because they’re so low to start with and they’re expected to jump up, but we [also] have a much more transient population.”

In general, many teachers are frustrated with there being “too much” – too much testing and too much pressure or weight put on one test. At the same time, they see standardized testing as an inconvenience that is a necessary part of education now. And they do look to these tests as an indicator for self-evaluation in their teaching. The comments in Table 3.5 show some examples of this internalized pressure teachers have toward the test results.

Table 3.5
Teachers’ Internalized Pressures in Relation to MAP Results

As a teacher, I see the need for them, because I need to see if what I’m doing is working.
It frustrates me . . . I never know if I really prepared them or not.
And doing all these things without having your own ego wounds . . . I look at it . . . first to reflect, as a reflective piece on my teaching and where I thought I should be, where the data indicates I was.
It’s just when I compare them to other classes, it’s a little depressing.
I was really anxious to see if my kids gained.

Teachers should have a vested interest in the success of their students and, certainly, seeing gains in student achievement is an easy (although not necessary accurate) way to measure one’s effectiveness as a teacher. However, the teachers were not directly asked about the

pressures of the MAP testing on them. Instead, they were asked how they use the results from the MAP testing. Therefore, the fact that their responses reveal this personal, internalized pressure is all the more notable; they actively view the use of this test as a personal evaluation, often even before discussing how it is used for the purposes of student evaluation. As stated in Chapter Two (and, specifically in Table 2.5), ninety-one percent of the teachers believe that the administration sees the MAP testing as a reflection on the quality of teachers' instruction. It seems that the teachers are internalizing the pressures that they are receiving from the administration (and other sources). Since they believe that the administrators use this test to evaluate the quality of their instruction, the teachers also adopt the belief that this test is a valid indicator of their own quality as an instructor. Interestingly, only one teacher showed an awareness of this assimilation, both explicitly recognizing and rejecting this idea in her interview – after admitting to previously falling into this pattern:

I'm kind of out of that now. First, I did look at it that way . . . And maybe that's a maturing a little bit as a teacher. It's not always looking at the low stuff as punitive or looking at the high stuff as "Oh, I did so well," because sometimes kids teach themselves. You might not have done anything . . . They just got it . . . I know that the lower [the students'] were at the beginning of the year, the more you want to see at the end. I would like to see that, but quite honestly, if I don't, I'm not going to fall apart.

The most notable way that many teachers showed their frustration and pressures with testing was in how they began the testing with the students. Table 3.6 shows examples of the teacher comments to students as they entered the computer room to begin MAP testing in the Spring. In addition to the chaos of having students find their seats at the computers and follow instructions for how to log on, the teachers were often very short with their students (in ways that contrasted with their usual rapport with students as seen in observations in their classroom lessons).

Table 3.6

Examples of Teacher Comments to Students as They Begin MAP Testing

Grade 3	“All conversation needs to stop right now.” Then, she leaves the room and has the computer teacher take over, who says, “Use your common sense. Use what your teacher has taught you in the classroom because I cannot help you on your test.”
	“I’ve already suspended two today, I’m on a roll . . . Stop – I didn’t tell you to do anything so that’s your problem right there.”
Grade 4	“Now let me see if you brought the things I told you to bring with you.”
	A student asks, “Can we use this paper?” and the teacher responds, “Yes, but you didn’t follow any of the rules, didn’t raise your hand, so no recess today.”
Grade 5	The classroom teacher says: “Stop asking questions and just listen to instructions.” Then the computer teacher states, “You have 52 questions and we have . . . we’re running late today . . . Go ahead and get started.”
	“Well, if you remember your score, try to do better. Do your best and try to increase your score . . . 52 questions. Let’s stay positive.”
	“You have 52 questions. Do your best. Do your best. I know you can. If you set a goal, your goal is to increase. You took this test in the Fall, remember?”
	“There are 52 math questions. If you already know what your Fall score is, try to make gains. I encourage you not to get stuck on one problem for too long. Use those strategies that . . . you have been learning all year long.”

As seen in the table, the comments to the third and fourth grade students as they began testing were more negative and pressure-filled than those to the fifth-grade students. This may be because the fifth-grade students have had a longer history with testing and, therefore, the teachers feel that they know what is expected of them. Even these teachers, who seemed less frustrated at the moment the test was beginning, still clearly cared about encouraging their students to score well and make gains. These teachers were not, therefore, necessarily experiencing any less pressure about their students scoring well, but rather, did a better job of not revealing their level of stress to the students.

Observational evidence also showed the concern that teachers have over the circumstances surrounding the testing. In particular, the Spring administration of the MAP

testing falls at a time when there are many distractions, such as field trips. The school chosen for this case study encourages the teachers to plan field trips after the state tests so that instructional time leading up to the state tests is not “wasted” on out-of-class experiences. Therefore, when the MAP testing was happening, one fourth-grade teacher commented, “We had a field trip Friday and Monday and more tomorrow and Thursday. They just aren’t focused. They’re rushing through. They had a hard time settling down this morning to even get started.” Another fourth-grade teacher commented on the student complaints that she had been receiving during their testing: “It’s too hot, I don’t understand this one, etc.” In general, the teachers seemed to agree that their students were more distracted at this time than in previous administrations of the MAP test.

The results of the tests, however, were clearly the focus for the teachers and the students. Of course, the teachers, like the students, get very excited when students make large gains and score well. Their surprise and excitement is both very visible (smiles, high fives, thumbs up signs) and vocal (“Wow!” or “What? Cool!” or, simply, “Good Job.”). When teachers would come to pick up their students from testing, the first thing that many of them did was ask the students to show her their scores. For example, a third grade teacher walked into the room and announced, “If you’re done, you should be reading. Those of you who have scores, let me see them.” A fourth-grade teacher commented to me while I was observing her class testing, “This is not a good day. None of them are making their goals.” This same teacher also told me that the school administration told her that all of her students needed to have an RIT score of 211 or higher for fourth-grade and that she needed to show 17 point gains for all of her students from Fall to Spring. A second fourth-grade teacher commented to me, “This is even worse than I thought it would be.”

How do administrators perceive the effects of MAP testing on the teachers?

The administrators are aware that the commitment and expectations placed upon the teachers by this additional testing is a source of some stress. However, they tend to see this as a positive part of the testing. As an instructional coach commented, “I think it depends on a teacher’s personality – whether that increased level of concern is productive or not. Some teachers just get in a tizzy and don’t change their instruction. And other teachers don’t get in a tizzy and do change their instruction.” The principal also explained, “It’s more work than what it would have been if the tests were not in place, but the more work gives them better results.” Along with this awareness of the teacher’s concerns and additional workload, the administrators believe the teachers see the greater benefits of the MAP testing. Table 3.7 shows a quotation from each administrator interview demonstrating this perception that the teachers are happy with the MAP implementation.

Table 3.7

Administrator Perceptions of Teacher Reactions to the MAP

[The teachers] are good sports about it. I think they appreciate having some kind of objective measure to make sure they’re making some growth.
At first, they were excited about the information. But the more they learned how to use that information and how to select which data they need to be looking at in order to . . . have effective instruction in the classroom, I think the more they get involved with that, the more exciting it is . . . I wish I had had a test like that [when I was teaching] . . . It’s exciting for teachers to sit down, pinpoint what the needs are, plan instruction.
Some of them do see it as beneficial in terms of using the information, being able to determine whether their children are making progress or not.

The administrators, therefore, seem to be unaware of the teacher complaints and frustrations or simply dismiss these responses as teachers just being unwilling to work harder at their

instructional practices. Interestingly, none of the administrators expressed concerns about the amount of data or the form of the data (i.e. the RIT score format) but only commented that the teachers are getting better at using the data over time.

What are the perceived effects of MAP testing on the students?

Qualitative Results. In their interviews, the teachers responded to a question about how the students react to the MAP testing. The responses varied from describing the students in various states, from “confident” to “burnt out.” However, only one teacher reported all positive reactions, explaining that her students “were pretty confident and they looked forward to . . . getting in there [and] taking the test.” The other teachers all reported at least some, if not all, negative reactions from the students, especially by the Spring assessment, when the students would take the MAP test for the third time that year, after just having completed the state’s KCA assessment. Over half of the teachers used the phrase, “another test” in describing their students’ reactions, with many of the teachers noting that the students are just over-tested by the Spring and tend to “shut down.” A majority of the teachers spoke of their students being “frustrated,” “burnt out”, “overwhelmed,” or “stressed.” A typical explanation from the teachers is exemplified by this statement:

My children, honestly, they just don’t care anymore. They’re burnt out. We test and test and test. And they’re just burnt out. And I reminded them that we have MAP testing towards the end of the year. It’s like, ‘So?’ And so I really don’t know how well they’re going to do because they’re tired. We put an awful lot of pressure on them . . . and they are really tired.

Despite the burn out, students do not just go through the motions of the test without being affected by it. They become more frustrated by the amount of testing, the material, the scores, and the external pressures. In almost every class, a student asked before or during the testing,

“How many questions is it?” or when they were told that there would be 52 questions, responded with impatience. One student exclaimed, “52? That’s more than last time!”

Specifically, students are frustrated by the material over which they are tested. Since the MAP test is designed to adjust to the level of each student, all students should be answering about half of the questions presented to them correctly. Students then get confused as to why they are being asked about mathematics topics that they have never been exposed to in class. One teacher reports that “Some of [the students] say . . . ‘I don’t know that stuff’ and ‘Why are you giving us a test when we don’t know stuff?’” A third-grade girl was an hour into the test and was only on question 16 (out of 52). She was working hard, tried to ask questions several times, but was frustrated with the difficulty of the questions in front of her. She gave a clear impression that she wanted to do well – so she did not just want to guess and move on – and yet she did not know the material on which she was being questioned and was even more frustrated that, when she asked for help, there was nothing the teacher could do to help her. She was the last student in her class to finish the test and I volunteered to walk her back to her classroom where the other students had been for awhile. As we walked, she told me that the test is hard “because it asks stuff from everything. I only improved two points . . . but I was supposed to improve more than that . . . I did try, especially on the fractions. And I couldn’t remember mean.” A fourth grade student made a similar comment to me while testing, “I don’t know how many inches are in a yard. We didn’t learn anything about that this year.”

While observing students testing, I saw some of the mathematics problems that MAP testing presented to the students and, like the students and the teachers, I was surprised by some of the questions asked. I did not expect to see rote memorization types of questions such as the number of feet in a mile. Nor did I expect that students would be asked about sampling methods

for a survey. Even in terms of the computational tasks asked of students, I was surprised to see elementary students presented with questions containing x and y input/output tables. (Of course, this would only be presented to a student who was answering easier questions correctly.)

I also observed ways that the questions and possible responses could misrepresent a students' true mathematics abilities. For example, a kindergartner was testing while I was still observing some older students in the computer lab. On this kindergartner's addition and subtraction questions, she was supposed to represent her two digit answers by sliding the correct digits into the two boxes. This student was computing the correct answer but reversing the digits, e.g. putting "91" where she intended to put "19." MAP results, however, will not show the teacher that this was the error. Instead, this student's results will simply show that she is low in two-digit computation when, in fact, her computation is fine but her problem is with number representation.

In every class, students had questions while they were testing. And questions were offered by almost all of the students, not just by students who were struggling with the language or those frustrated with the problems. Unless they were technical questions (for example, about logging into the test or moving onto the next question), the teachers were not allowed to help. Some students asked about the mathematics concepts, generally ("I forgot how to do this problem" or "I don't get this") or specifically ("What do these lines mean?" while pointing to a negative symbol). Other students were not clear about how to use their scratch paper – writing down all of their questions and answers on the paper or trying to figure out how to show all of their work on questions to which they had already found the answer. I noticed a third-grade student who was only on question three after 45-minutes of testing, yet she seemed to be really concentrating. I told her to make sure that, once she has her answer, she clicked "go on" to get

the next question. She answered, “I’m thinking about how to show my work.” And so I explained that she did not have to show her work for every problem and to keep going. Five minutes later, she was up to question six. Another student complained to me, “This test only gives me a calculator on questions when I don’t need it.”

One of the most common questions students have is on vocabulary and reading. For example, one student asked me: “What’s this word?” and pointed to the beginning of a word problem that said, “Yori is . . .” In this case, the word was just the name of the person in the word problem, but the student thought it was a key mathematics word that was necessary to the problem. Some students asked about words that were mathematics terms necessary to the problem, such as “average” and “horizontal,” or terms in a word problem that were important to understanding the concept, such as the word “poured” in a word problem about capacity. While many students had vocabulary issues, this was especially problematic for the English language learners. A fourth-grade teacher who had most of the ESL students in her classroom told me, “It doesn’t seem fair for the ESL kids to have to read it themselves,” but she explained that their instructional coach had told them that they could not even read specific words out loud to the students.

When students were confused about what the question was asking, they were more likely to get stuck and not move onto the next question. Often the teachers would recognize this and prompt them to move on, saying, for example, “Come on . . . don’t get stuck on a question. Just pick your best guess and move on.” Other students, however, would just click through the test without attempting problems. Teachers also prompted students in these cases to slow down, asking “Why are you going so fast? You have scratch paper and you haven’t even used it?” or telling a student, “Stop guessing and work out the problems on paper.” Students also seemed to

be influenced by how fast their neighbors were going on the test – often looking to see what number question their neighbor was on and adjusting their pace accordingly or even commenting to the teacher that he/she was farther behind. One student called me over when his neighbor finished the test about 45-minutes into the testing time while he still had almost twenty questions left; I assured him that he was doing fine.

Two of the teachers talked about how they try to get their students “psyched up” and ready to test, including one teacher who uses toys as prizes when students raise their scores. Many teachers reported that students want to know their scores and if they are improving. As a general guideline at this school, the teachers are supposed to give students a goal score as they go into the MAP test – many carry them on a post-it note and place these on the computer screen as they test. In one class, the teacher sent the students with post-it notes that also included a blank where there were to fill in their actual score when they were done. For some, this creates a very personalized pressure in itself. One fourth-grade girl that I observed broke down in tears when her assessment score popped up on the screen as she finished her MAP test. She cried for about an hour, quietly in her chair with her head on the desk. The teacher explained to me that this student knew that she needed a higher score to qualify for the gifted program, so she was very disappointed when she did not get that score. An instructional coach recounted a similar scenario about two first-grade students during their MAP testing experience: “I had a boy and a girl last week that just freaked out and cried.”

But these scores, themselves, also cause confusion for the students. Because they are reported as RIT scores (and not percentages), the students do not have a point of reference to know what their score means. One teacher said, “They do want to know their scores but they don’t understand them because it’s not ‘seventy percent,’ it’s ‘204.’ They don’t line up, so they

don't get it. 'OK, well I scored 157.' They think that's 157 percent. They don't understand that that's not that good." This also means that the students do not understand when they have done well. For this reason, during testing, students often waited for the teachers to tell them that their score was good before they got excited. One student, who had his goal on a post-it note as a 203, mistakenly thought that was his previous test score. So, when he saw that his score was a 205, he thought he had not improved enough. He disappointedly told his teacher that he did not make his goal, but, when the teacher explained that he did – he had improved seven points over his last test – he became very excited, smiling, and bouncing in his chair. Another student made her goal score exactly and mistakenly assumed that this was a perfect score, asking her teacher, "So I got them all right?" The students can also be confused by the inconsistency between their grades in class and their test scores. This problem was brought up by a teacher in her interview who explained, "It could be negative because they study a lot and they do really great work and they get positive grades . . . but those are not consistent with their scores on the MAP test."

For some students, the teachers see the need to try to create reasons for them to internalize some pressure toward their MAP scores. One teacher explained to her students that "these scores are going with you to the next grade" while another was even more direct: "I kind of threaten them I said, 'If you don't want to take two hours of math next year, you really need to do well on the MAP because that decides whether you take two hours of math or one hour of math.'" When the students feel successful on the test, they are excited, even if they do not understand why the score is important. Many students seemed motivated solely for the feeling of accomplishment. For those that met their goal scores, it was a time of celebration – students would often gasp or jump out of their chairs raising their hands as soon as they saw their scores. But, for those that did not meet goal, they expressed their disappointment in themselves

– some tears, often just sitting quietly or expressing their anger that the test asked about things that they had never been taught.

The administrators also reported concerns about the stress of additional testing on the students. Both of the instructional coaches talked about the kids being “tired” and “fatigued” by all of the testing. Similar to the teacher concerns, one instructional coach also called the MAP assessment “just another test” that the students find “dry.” Other concerns voiced by the administration were that the testing can get in the way of more authentic education. One administrator explained:

The sense of an education gets lost in test prep. And that worries me because we may have kids that are quite competent at passing test, but they don’t have an education. They can’t think . . . Sometimes in the mania to be able to measure that we’re educating the kids, we’re not educating the kids.

For these reasons, the administration seemed to express sympathy and, perhaps, even some guilt for all of the testing in which the students must participate. An instructional coach explained, “They’re amazingly patient with all the testing we do with them. I don’t know why they do it . . . but most kids are good sports about it . . . [but] some kids come in feeling a great deal of pressure.” The principal echoed this concern: “I’m just hoping that using so many tests for students does not take the focus off of the fact that students are individuals and that they have other needs than what we obtain from testing.”

Quantitative Results. The teacher survey results regarding the effects of MAP testing on the students corroborate the qualitative results. According to the teacher surveys, just over half of the teachers believe that their students try to score well on the MAP assessment. (These results are presented in Table 3.8). This same number of teachers also believes that the students feel extreme anxiety toward this test, with two of the teachers strongly agreeing with this statement.

Table 3.8
Teachers' Opinions of Student Responses to MAP

Survey Statement about the NWEA-MAP Assessment	Percent Agreeing
The majority of my students try their best on the NWEA-MAP assessment	55**
Many students are extremely anxious about taking the NWEA-MAP assessment.	55**

**Two strongly agree.

The qualitative results help to explain why almost half of the teachers disagreed with these statements. Primarily, this would seem to be because the teachers perceive that the students stop caring about the test. This is particularly true by the end of the school year when the teachers report that the students are over-tested and “burnt out.”

Analysis

There is an environment of pressure that is created around assessment at the school. The MAP testing is one component of the pressure. And this pressure is felt by all – administrators, teachers, and students. The administrators are under pressure to demonstrate AYP on the state tests, to show improvement in student scores, and to justify their instructional decisions with data. This pressure is passed onto the teachers, who feel that they are being held accountable for student test scores, the results of which they feel they can do little to truly affect (as explained in terms of their outcomes expectancies, described in Chapter Two). The teachers feel that pressure coming from multiple sources, but most especially from themselves; they have internalized this pressure for their students to do well and, therefore, they tend to see the student results on MAP (and other standardized assessments) as a valid indicator of their own teaching quality. The students also feel the pressures of multiple standardized tests, of the weight placed on their scores for decisions about their education, of perceived competition with their peers, and of their internalized desire to improve their scores and to demonstrate their improvement to their teachers.

Student Frustrations with MAP Testing

As the teachers and administrators recognize, the students are overwhelmed with the amount of testing as indicated by their comments: “Another test?” or “Are we taking another state assessment?” They are also intimidated by the length of the MAP assessment (52 questions) and by the types of questions on the MAP. The questions on the MAP are formatted differently (in terms of the language and vocabulary used and the types of questions asked) than the KCAs and, by extension, the practice KCAs that they are most used to taking. Additionally, the MAP intentionally presents them with questions that are beyond their mathematics ability in order to best gauge their current level of mathematics ability; but students, instead, see this as “unfair” because it is not material that they have been taught but for which they know they are responsible on the test. The vocabulary, especially for the English language learners, presents its own challenges and frustrations, particularly when students are told that the teachers cannot help them with words that they do not know. Finally, the students are confused by the RIT scores that they are given; since they are not familiar with these scores in any other context outside of MAP testing, they do not know whether or not a particular score is “good” or “bad,” nor do they understand how to relate it to their other mathematics scores or grades.

Teacher Concern for Immediate and Long-Term Effects for Students

The teachers realize that the students are under immense pressure to do well on all of the assessments that they are asked to take throughout the year. The teachers see this pressure on their students and are concerned about the effects of this continual testing on them. They want to motivate their students to do well but are concerned that the kids feel burned out. As the teachers truly desire to do the best job that they can for their students, they become very conflicted. Good test scores are good for their students when the school is recognized for

meeting AYP, for example. However, the teachers are also concerned about the immediate effects on the students and do not want to inflict all of this testing and pressure on them.

A conflict seems to exist between protecting the students from testing pressures while motivating them to do well. Pressure is intentionally passed from the teachers to the students in ways that are meant to encourage them to work hard and do well: the post-it notes with goal scores, physical rewards for scoring well, and verbal praise and excited gesturing when meeting or surpassing goal scores. Unintentionally, teachers also convey pressure in the ways that they bring up testing in their regular mathematics lessons or even threaten students that the scores will affect their future mathematics classes. Additionally, the students are affected by their teachers' stress as seen in how many of the teachers began the MAP testing with their students – yelling, taking away recess, threatening suspensions, and making other negative comments toward the students before the test began. Yet, despite all of this, the teachers do make an effort to shield the students from getting “burnt out” or from stressing about their results to the point of breaking into tears at the results (as some do).

Teacher Concern Over Standardized Testing

The teachers also have concerns about standardized testing, in general, due to the bias that can be present in the testing and the inability of the tests to give a complete picture of a student. Some teachers were concerned about students who do not test well or who have a “bad” day when taking the test. Other teachers were frustrated with the amount of standardized testing and questioned whether students can take the test seriously when they are so over-tested. Therefore, the teachers question the reliability of the scores and are frustrated that so much weight is placed on scores that they do not trust are accurate representations of their students' mathematics abilities.

Additionally, the teachers are confused about what to do with the additional data they receive. As discussed in Chapter Two, they believe that additional data are valuable. However, they expressed concern about the lack of time that they have to truly analyze the data or use these data effectively in their lesson planning. This is in addition to the general concerns about the amount of time spent in testing itself. The administrators agree with the teachers about this concern; the amount of time spent in testing along with the number of standardized tests administered (in particular, the MAP, KCA, and often weekly practice KCAs) is overwhelming for the students and teachers.

Perhaps most importantly, the teachers are concerned about how student performance on standardized testing reflects on their teaching. The teachers see this pressure coming from the district and school administrators and are frustrated by the ways that the results are often used to evaluate their teaching rather than evaluate their students' needs. But, interestingly, the teachers feel the greatest pressure coming from themselves – they have internalized the pressure and see the students' results as at least one way that they receive feedback on their own achievement as teachers. Even as the teachers are frustrated by the weight placed on the MAP results by the administration, they tacitly accept the measure as a valid means for evaluating their teaching.

Administrator Misunderstanding of Effects on Teachers

The administrators realize that these pressures and concerns exist but are quicker to dismiss them than are the teachers. The instructional coaches, for example, did mention over-testing and student burn-out in their interviews. However, the instructional coaches and the principal see the pressure on teachers as justified and seem to believe it only poses a problem for teachers who are not willing to work hard to do better. This disconnect from the experience of the teachers – who feel pressure to do well while sensing little control over the results – adds to

the frustration of the teachers. The teachers not only feel that they have to protect their students but also react out of self-defense to protect their classroom and their instructional decisions from judgment. The teachers want to improve and welcome data that help them to improve, but the MAP data do little to help them improve their instruction. Therefore, as was also seen in Chapter Two, MAP results seem to be used more for evaluative judgments of their teaching rather than as constructive feedback on how to improve their teaching.

Effects on Instruction

The difference in instruction seen after MAP testing compared to lessons observed prior to the Spring administration of the MAP test was dramatic. In some ways, this difference may be seen as a reason to support the use of MAP testing – the teachers were teaching mathematics lessons that included vocabulary development and higher-order thinking skills on grade-level appropriate topics leading up to the MAP testing. And, once the testing was over, and therefore the pressure was off, the quality of instruction was negatively affected, with some teachers no longer even teaching mathematics lessons. In part, of course, this is a common pattern for the end of the year in any school. But these observations were not done in the last week of schooling, and so one should expect instruction to still be occurring “as usual.” Therefore, one could argue that the presence of testing accounts for the better instruction and the lack of testing and accountability leads to ineffective instruction.

This conclusion, however, would ignore the bigger picture presented in this chapter. It is precisely because of the pressures and stresses leading up to testing that the teachers and the students “collapse” after testing. The “burn-out,” discussed by the administrators and the teachers in their interviews, is *caused* by the testing. And, therefore, the testing is at least as much the cause of the poor instruction afterward as it is the cause of the quality instruction prior

to testing. I would argue that, therefore, without the pressures, quality instruction would be more consistent. It would also be more purposeful since this dramatic difference in instruction demonstrates how the presence of excessive testing leads to a teaching-learning environment in which purposes and motivations become solely good test performance rather than the inherent (and more authentic) value of learning the material.

For the teachers, as well, the scores can easily become the goal rather than the means to planning and improving instruction or meeting the individualized needs of students identified by the testing. Teachers are left with little motivation (within the limited time that they already have) to deal with the dense and, often, confusing data they receive from MAP testing. However, obtaining and *using* these data are the stated reasons that the administration values the continued implementation of MAP. Therefore, Chapter Four will examine what teachers do with the data they receive from MAP testing; in particular, the chapter will address how teachers (along with the administration) analyze, interpret, and apply these results. Additionally, in Chapter Four, a quantitative study examines whether information provided through MAP data is truly adding to the teachers' understanding of their students' mathematics abilities and providing them new information as to the best ways to support mathematics learning for their students.

Chapter Four

Is This Additional Testing Providing New Information to Teachers?

As seen in Chapter Two, the teachers reported using other standardized testing data, such as the KCAs, more often and more directly in their instruction than the MAP. This chapter will further explore how the teachers at the school presented in this case study are using the data they receive from the MAP test, specifically how they approach the analysis of data as they receive them. Beyond just the use of data, however, this chapter will detail the teachers' understanding of the MAP data, since they cannot use these data effectively if they do not correctly understand and interpret what they receive. This is an especially relevant question with the MAP results, since many of the scores are reported in the form of Rasch Unit (RIT) scores, a type of scoring not used on any other standardized assessments administered at this school. Moreover, this chapter examines a need for this additional data based on whether or not teachers can predict how their students will score on the assessment since, if teachers can accurately predict their students' scores, then the data are not providing them with new information or a better understanding of the needs of their students. Therefore, this chapter addresses the question of whether teachers can accurately predict how students will score and, therefore, if they already know the information provided by MAP results. A related question is, then, do teachers know how to interpret these RIT data that they receive from MAP testing? And, more generally, when teachers receive the MAP results, how do they interpret and analyze these data?

Data

Both qualitative and quantitative data were used to address the question of whether additional testing is providing new information to teachers. The quantitative data comprise the primary means of answering this question. These data include teacher score predictions, student

test scores, and teacher survey data. The qualitative data – in the form of teacher and administrator interviews and classroom observations – help enhance the understanding of the quantitative data presented.

At the school where this study took place, the Measures of Academic Progress (MAP) test is administered three times each year. While this study focuses primarily on the Spring 2009 administration of the MAP, it references scores from the Fall 2008 administration in order to look at student growth over the school year. The Winter 2009 scores were also available but were not used. The data sheets provided through the online NWEA-MAP website can be viewed by class or by individual student. All of the students' previous scores on the MAP, from any time they have taken the test while they have been enrolled in the district, can be accessed by the teachers and administrators in this way. Students who transferred into the school during this study and had a Fall 2008 score available from their previous school were included in the data collected for this study.

Prior to the Spring 2009 administration of the MAP test, teachers filled out a form with their predictions of their students' overall RIT scores as well as their predictions for the score ranges (low, average, or high) for their students in the four sub-categories of the mathematics assessment (numbers and computation, algebra, geometry, and data). These score ranges match the way that the MAP assessment reports students' scores to teachers. The teachers were given the students' names and Fall 2008 scores but were not allowed to view any additional information, such as the students' grades or Winter 2009 MAP scores. A sample of this form is shown below in Figure 4.1. The assumption underlying this portion of the study is that teachers who predict student scores well must already have the information required to accurately assess

their students' mathematics achievements even without the data provided by this additional assessment.

Figure 4.1

Sample Teacher Prediction Instrument

Ms. Jones 3rd Grade		Predictions for Spring 2009 <small>NOTE: LO = 33rd percentile or lower; AV = between 33rd and 66th percentile; HI = at or above 66th percentile</small>				
Student Name	Fall 2008 RIT Score	Spring 2009 RIT Score	Number & Computation	Algebra	Geometry	Data
Johnson, Joe	153		LO AV HI	LO AV HI	LO AV HI	LO AV HI
Smith, Sally	184		LO AV HI	LO AV HI	LO AV HI	LO AV HI

While the usefulness of the MAP data is first analyzed in terms of the teachers' score predictions, other data from the teacher surveys provided insight into the teachers' ability to accurately understand and interpret the RIT data. If the teachers lack an understanding of the data provided, it is unlikely that they are learning useful information based on the data.

Therefore, the teachers in this study completed a survey instrument that included a set of ten true or false questions that focused on RIT scores to address the question of whether teachers are receiving new information from the MAP testing. (This survey also included questions for other parts of this study, such as demographics, teaching background, mathematics teacher self-efficacy, and perceptions of the MAP testing.)

In the three weeks following the completion of the survey, all eleven third, fourth, and fifth grade teachers also participated in an interview process about the NWEA-MAP assessment. The interviews focused on five main areas, as described in Chapter One and seen in the interview protocol provided in Appendix B. For the portion of the study discussed in this chapter, responses to the fifth part, the teachers' overall views of standardized testing and the MAP

assessment, are the most beneficial. However, relevant comments from other portions of the interviews are also included in the results and analysis for this chapter. An examination of these themes, along with comments that diverged from the pattern, enhances the understanding of how teachers approach the NWEA-MAP data and the ways in which they interpret, discuss, and apply these data. Similarly, relevant portions of the administrator interviews (which were conducted during the same few weeks as the teacher interviews and the protocol of which is provided in Appendix A) are also included in this chapter. Additionally, classroom observations (from before, during, and after the Spring 2009 administration of the MAP assessment) help to corroborate the teachers' self-reflections on these issues.

Methods

Quantitative Methods

Data were collected on 234 students and 11 teachers. Due to students moving both in and out of the school (and district) during the year, or missing the testing window through absences, completed data for analysis were available for only 188 students. For example, the data set included 228 student scores for Fall 2008 but only 188 student scores for Spring 2009. Thus, for any statistics related to student results on the Spring 2009 assessment, only 188 cases were available for the analysis.

Student data include RIT scores from the Fall 2008 and Spring 2009 NWEA-MAP assessments along with their Spring 2009 ranges (low, medium, or high) in each of the four subcategories of the mathematics test. One variable, student growth, was calculated based on the difference between the Spring 2009 and Fall 2008 scores. These numbers were then analyzed alongside the teacher predictions of student RIT scores for the Spring 2009 assessment, both in terms of the RIT score, itself, and in terms of predicted growth, i.e. the difference between the

Spring 2009 predicted score and the students' actual Fall 2008 scores. Another variable, error in predicted growth, was defined as the absolute value in the difference between actual student growth (i.e. difference between Spring 2009 and Fall 2008 RIT scores) and the teacher predicted growth.

The analysis began by seeking correlations between predicted scores, actual scores, predicted growth, actual growth, and the error in predicted growth. Next, correlations between predictions and actual achievement of students in each of the four sub-categories of the mathematics test (numbers and computation, algebra, geometry, and data) were calculated. Additionally, paired-sample t-tests were used to examine the difference of means between predicted growth and actual growth. In this case, no evidence of a statistically significant difference would indicate good predictions. (Refer to Table 4.1.) These t-tests were first done on the whole sample ($n=188$) and then on the class of each individual teacher. For the individual class data, the sample size was greater than or equal to fifteen in all but one case, allowing for inferences to be made based on the t-test.

Since the sample size of teachers ($n=11$) was so small, the teacher survey data only allowed for a descriptive analysis. For the survey items analyzed in this portion of the study, the data are presented as a percentage. Table 4.5 includes the percent of teachers agreeing with various statements from the survey related to MAP results and MAP data. The statements included in this table were adapted from a national survey used in a study by Pedulla, et al. (2003). Please refer to the Methods section in Chapter Two for a further description of how items were selected and adapted for use in this study. In Table 4.6, the percents presented indicate the number of teachers correctly answering a true or false statement (for ten statements) about RIT scores and MAP testing. These ten true or false items were developed based on

NWEA documents about using and understanding MAP (Northwest Evaluation Association, 2007, 2008a, 2008b, & 2009a).

Qualitative Methods

While the majority of this part of the study is based on the quantitative data, qualitative data are included to support and broaden the understanding of the quantitative findings. Therefore, relevant quotations from the teacher and administrator interviews are included in this chapter. The Methods section in Chapter Two explains, in detail, how the interview data were collected and analyzed. Similarly, observational data from the MAP testing and from my interactions with the teachers in their classrooms supplement the interview data. The methods of collecting and analyzing the observational data are described in the Methods section of Chapter Three.

Results

Quantitative Results: Score Predictions

Based on the score predictions, teachers expected an average student growth of 6.8 units (on the RIT scale) between their Fall 2008 and Spring 2009 scores. Students actually had a greater average gain than the teachers predicted; the average gain was 11.5 units. Therefore, the teachers under-predicted student achievement. Teachers' errors in prediction averaged 7.0 units. Table 4.1 shows a comparison of this overall result with the means by grade level.

Table 4.1

Means of Predicted Growth, Actual Growth, and Error in the Predictions of RIT Scores

	Mean Predicted Growth	Mean Actual Growth	Mean Error in Predicted Growth
All	6.8	11.5	7.0
3 rd Grade	6.9	13.8	8.1
4 th Grade	6.8	9.8	6.3
5 th Grade	6.7	10.8	6.6

The error in predicted growth was highest among third-grade teachers (8.1 units), while fourth and fifth-grade teachers had less error in their predictions (6.3 and 6.6 units, respectively).

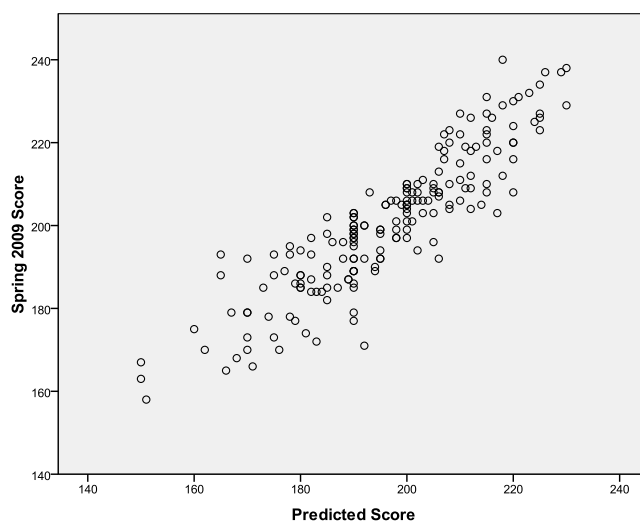
However, at all three grade levels, the under-prediction pattern held.

These results differ from those of Doherty and Conolly (1985) who found that teachers generally overestimated student mathematics performance. Rather, in this study, only 25% of the predictions were overestimations of student performance. The third-grade teachers' predictions were only higher than 18% of their students' scores. Fourth- and fifth-grade teachers had predictions higher than their students' actual performance in 30% and 27% of the cases, respectively.

The teachers' predictions were highly correlated with the students' actual scores ($r=0.90$, $p<0.01$). This high correlation is clearly seen in Figure 4.2, a scatter plot of the predicted RIT scores and the actual student scores. This scatter plot also suggests no strong outliers in the data.

Figure 4.2

Scatter Plot of Teacher Predictions and Student Actual Performance on Spring 2009 MAP Test



The predictions of achievement levels in each sub-category of the mathematics assessment (numbers and computation, algebra, geometry, and data) were less strongly

correlated with the students' actual performance, but these correlations were still strong and significant. As seen in Table 4.2, teachers better predicted student achievement in data and algebra than in geometry or numbers and computation.

Table 4.2

Correlations between Teacher Predictions and Student Achievement

RIT Score	0.90**
Data	0.68**
Algebra	0.63**
Geometry	0.55**
Numb & Comp	0.47**

**p<0.01

Unlike in the Pezdek, Berry, and Renno (2002) study where parents' discrepancy scores were correlated with the child's achievement, in this study there is virtually no correlation ($r=.08$) between the teachers' discrepancy scores (i.e. their error in predicted growth) and the students' actual scores. In other words, teachers did well at predicting the scores of students at all achievement-levels, not just the scores of the high-achieving students. However, teachers were better at predicting the achievement of students who demonstrated greater gains in their RIT scores during the year since the error in the predicted growth was significantly correlated with the students' actual growth ($r=0.46$, $p<0.01$). In other words, the results of this study do not show that teachers are better at predicting the scores of higher-achieving students but, rather, that they are better at predicting the scores of students who are making the greater gains in their achievement. These are, often, students who begin the year scoring low but then make greater progress during the academic year; in fact, many of these students would be the "bubble" students who are often targeted for test preparation so that they can be classified as "proficient" on the state-mandated tests.

Studies by Coladarci (1986) and Helmke and Schrader (1987) both found the accuracy of prediction varies by teacher and classroom. As one would expect, this was true for the teachers in this study, as well. While there was an overall statistically significant difference between the means of the predicted growth and actual growth (as well as the predicted and actual scores), the predictions of student growth of one (of the four) third-grade teachers, one (of the three) fourth-grade teachers, and two (of the four) fifth-grade teachers showed no statistically significant differences from the students' actual growth (see Table 4.3), indicating a high accuracy in their predictions.

Table 4.3

Paired-Sample T-test, Difference of Means of Predicted Growth and Actual Growth for Individual Classrooms

Teacher	t	n	Mean Actual Growth	SD for Actual Growth	Mean Predicted Growth	SD for Predicted Growth	Mean Error in Predicted Growth
ALL	-8.482**	188	11.49	7.60	6.78	5.66	7.01
A	0.259	17	14.82	7.83	15.29	4.00	4.59
B	-0.807	15	10.47	4.36	9.13	3.96	4.80
C	-1.271	17	9.41	8.55	6.94	2.78	6.65
D	-1.749	21	9.29	8.44	6.10	5.60	7.19
E	-2.228*	16	8.19	8.63	3.81	2.66	6.88
F	-2.438*	17	15.53	8.27	11.65	6.79	6.59
G	-2.492*	21	10.57	7.54	7.48	4.76	5.19
H	-4.842**	15	10.87	7.96	.69	3.59	9.80
I	-4.895**	17	9.00	6.09	2.18	1.91	7.76
J	-6.716**	13	15.77	3.79	7.29	2.76	8.23
K	-8.596**	16	14.00	5.61	3.69	2.91	10.31

*p<0.05. **p<0.01.

The difference between the mean predicted RIT scores and the mean actual RIT scores for a class differed from 0.95 units to 10.3 units. This demonstrates the differences in accuracy among individual teachers in predicting student achievement. However, while over half of the teachers' predictions and actual student growth did have a statistically significant difference in their means (indicating less accuracy in prediction) the standard error reported in the Spring 2009 RIT score reports for the students at this school was generally around three. This means that a student's score could legitimately indicate a seven unit range of scores. (Note in Table 4.3 that this matches the mean error in predicted growth for the teachers participating in this study.) In this light, even the largest difference of mean (10.31 units) does not seem that extreme.

Qualitative Results: Teacher Interviews and Observations of MAP Testing

The qualitative results allow for a closer look at the four teachers whose predictions in growth were not statistically significantly different than the actual growth of their students. Is there something in common among these teachers that makes them unique from the other teachers whose predictions were less accurate? The data revealed no patterns in terms of years of teaching, years of teaching at this particular school, year of obtaining a bachelor's degree, type of institution where the degree was attained, or areas of certification. Similarly, classroom observations of lessons and, in particular, student engagement levels and levels of cognitive activity found in the lessons, were not unique among this group of teachers who predicted student growth on the MAP so well. Even the quantitative data from the teacher surveys, such as the teacher efficacy scores and outcome expectancies, did not highlight a difference between these four teachers and their colleagues. Since all of these factors did not help to explain this difference and, most importantly, in order to protect the anonymity of the teachers, these data on individual teachers are purposely excluded from the results in this chapter.

However, two areas in the qualitative data did emerge as differences for the four teachers whose predictions were most accurate. The first difference was the use of post-it notes for the goal scores during MAP testing. The four teachers with the most accurate predictions (based on the paired sample t-tests) were the same four teachers who gave their students the post-it notes to place on or next to their computer during MAP testing. Therefore, while many teachers mentioned that their students knew their previous MAP scores or their goal scores as they went in to take the Spring administration of the MAP, the students of the four teachers of interest to this finding actually had their previous score or goal score with them and in front of them during the MAP test.

The teacher interviews provided the second difference between these four teachers and their colleagues: the teachers' views on standardized testing. Table 4.4 includes excerpts from all of the teacher interviews where the teachers were responding to the question, "How do you feel about standardized testing, in general?" Please note that the teacher labels (A, B, C, etc.) match up with those in Table 4.3, and that the first four teachers are bolded in both tables to indicate that they are the four teachers of interest who had the accurate predictions of student growth on the MAP (i.e. no statistically significant difference between their predicted growth and the students' actual growth between the Fall and Spring administrations of MAP). As seen in Table 4.4, these four teachers all hold primarily negative views of standardized testing, in general, while the remaining seven teachers tend to have more mixed or positive reactions to standardized testing. (The one exception to this is Teacher I, who also holds primarily negative views of standardized testing. Yet, even Teacher I's response may be categorized as more indifferent or unsure than overly negative.)

Table 4.4

Comments from Teacher Interviews on their Views of Standardized Testing

Teacher	Excerpts from Teacher Responses to the Interview Question: “How do you feel about standardized testing, in general?”
A	I think we do too much . . . I don’t like standardized tests. To me, it needs to be more authentic . . . not so much cram and test . . . But if you take data throughout the year authentically, that would be better than a standardized test where they fill in bubbles or click bubbles on a computer. I know it’s quick, it’s easy, it’s fast. I don’t think you’re getting true data.
B	When we’re expected to teach to the test, it’s a pain.
C	I hate it. . . . I just feel like we spend so much time and we walk that fine line of . . . teaching to the test . . . It just makes me sick.
D	It’s kind of a pain and I just wonder if there’s a better way of doing things.
E	In general, I believe it’s important to know where the students are . . . to be able to compare them to the nation . . . I believe standardized tests are important.
F	I think we need it. But I think there are better ways to show improvement than . . . the system they’ve come up with.
G	I have mixed feelings . . . I do think it helps to know where your kids are . . . [but] I think that there’s too much pressure involved in it.
H	It has its place.
I	I think it’s biased . . . I don’t think it always reflects what the kids are capable of and what they know.
J	I think it can be effective . . . I think it is a pretty good assessment of how the kids have learned during the year.
K	I think it’s necessary because I do believe . . . students in my class should be expected to know and should be exposed to the same amount of information [as] a . . . student [in another part of the country] . . . as much as humanly possible. . . I think they should be competitive.

Quantitative Results: Survey Questions

The teacher surveys reveal many positive perceptions of the teachers toward MAP testing. As reported by 91% of the third, fourth, and fifth grade teachers at the school, the teachers believe that the NWEA-MAP assessment provides useful information, and 64% indicate that these data included information that they would not otherwise know. (Please refer to Table 4.5.) More importantly, however, 91% report that they have been prepared to interpret the results from the assessment, with one teacher indicating strong agreement with this statement, but only 36% believe that the NWEA-MAP score reports are easy to interpret, with two teachers feeling strongly that interpretation of the MAP data is not easy.

Table 4.5

Teacher Agreement with Survey Statements about the NWEA-MAP Assessment

Survey Statement about the NWEA-MAP Assessment	Percent Agreeing
Professional development regarding the NWEA-MAP assessment has adequately prepared me to <i>interpret</i> the test results.	91*
The NWEA-MAP score reports provide useful information.	91
The NWEA-MAP assessment provides me with new information about my students that I would not otherwise know.	64
The NWEA-MAP score reports are easy to interpret.	36**

*One strongly agrees. **Two strongly disagree.

Despite the fact that teachers believe themselves to be prepared to interpret the MAP data, this difficulty in interpreting the data may lead to more confusion than the teachers are aware of or are willing to admit. The ten true or false items relating to the NWEA-MAP testing and RIT scores that were included on the teacher surveys help clarify where teachers are secure in their understanding and where they lack understanding about the data they receive from the MAP assessments. Table 4.6 lists some important facts about MAP testing and RIT scores along with the percent of the teachers demonstrating understanding of these key concepts. As seen in the table, the teachers demonstrated an overall understanding of percentile rankings, with 82% of the teachers answering this survey question correctly. Percentile rankings are the form of assessment results with which the teachers are most familiar since most external, standardized assessments tend to report their data with percentiles. While the MAP does not focus on the percentile rankings, teachers are given these data along with the raw RIT scores, RIT ranges, and other MAP-specific data. The teachers also demonstrated some proficiency with the basics of MAP testing and RIT data. Eighty-two percent know the score ranges for the MAP testing. Seventy-three percent understand that the RIT scale allows one to measure growth over time and that the RIT range should be used when making placement decisions. However, the teachers

were much less informed about issues regarding the consistency of student results, the meaning of the RIT scores, the typical growth to expect in RIT scores during the year, and the format for question selection for students.

Table 4.6

Percent of Teachers Demonstrating Understanding about NWEA-MAP Testing and RIT Scores

Facts about NWEA-MAP Testing and RIT Scores	% Demonstrating Understanding
The RIT Scale helps to measure growth over time.	73
The RIT Scale has the same meaning regardless of grade or age of the student.	45
When making a placement decision for a student, the score range (rather than the single RIT score) should be used to help make the determination.	73
NWEA-MAP assessments do not differ for each grade-level.	36
Students achieving typical growth in RIT scores will remain at about the same percentile score.	27
If a student retook the NWEA-MAP test about the same time, the student's score would fall within the same RIT range 68% of the time.	9
The percentile ranking compares students to a nationwide norm sample, not to their classmates.	82
All students will only answer about half of the questions presented to them correctly.	27
Reasonable increases in RIT scores differ by students and by grade level. The increases can be predicted using a target growth chart.	27
RIT scores range from about 100 to 300.	82

The most extreme result from this portion of the study was that only nine percent of the teachers understood the reliability in RIT scores. NWEA explains that students who re-take a MAP test should obtain a score within their RIT range “most” of the time; the NWEA clarifies that “most” refers to “68% of the time” (Northwest Evaluation Association, 2007). On the survey, all but one teacher believed that, if a student retook the test, his/her score would fall within the initial test's RIT range 90% or more of the time. This raises concerns about the

amount of trust that the teachers place in the consistency of these scores – a trust that goes beyond what NWEA actually claims to provide with these assessments.

Despite these results, a majority of the teachers believe that they are prepared to administer these tests (64 percent), and a smaller majority (55 percent) believe that they are prepared to apply the results from the MAP test. Teacher survey results from two items on professional development regarding the MAP assessment are shown in Table 4.7.

Table 4.7

Teacher Survey Questions on Professional Development Preparedness for the MAP Assessment

Survey Statement about the NWEA-MAP Assessment	Percent Agreeing
Professional development regarding the NWEA-MAP assessment has adequately prepared me to <i>apply</i> the results to my teaching.	64
Professional development regarding the NWEA-MAP assessment has adequately prepared me to <i>administer</i> the test.	55*

*One strongly agrees.

The district may believe that the teachers are adequately prepared to administer and use the MAP testing based on the teachers' own admission of their preparedness. However, the teachers do not seem to realize how much misinformation they hold about the MAP assessment and data.

Qualitative Results: Teacher and Administrator Interviews

At the same time, many of the teachers do find the RIT data to be “confusing” and “ambiguous.” Table 4.8 lists some of the teacher comments about RIT scores. Overall, the teachers seem to express some frustration with the scores and a general acknowledgement of their need for more clarity on the meaning of these scores. The six teacher comments in the table are representative of the general attitude of the teachers (and even, at times, the administrators) toward the RIT scores. As was described in Chapter Three, students also demonstrate a similar confusion with these scores. Consequently, since the teachers do not understand the results well, they cannot adequately explain to their students what these scores mean or how they are to interpret them.

Table 4.8

Teacher Comments about RIT Scores

Yeah, well, I wish those numbers meant more . . . I don't understand what that is, when I ask what is that specific one hundred and seven [or] one hundred and twelve number computation score and how that fits in with the RIT score . . . I don't know what number.
I understand RIT scores, but I don't think I fully understand, like, the number versus the scales. I don't fully understand that.
I mean, I see the numbers there, and RIT score, and all that, but it doesn't really mean anything until I actually, you know, can see what they're doing.
I have to be reminded [about RIT scores] and I am not used to them yet . . .
Sometimes certain aspects, like in the range . . . I am not always sure exactly, what the range means, you know, and what grade levels they are, I always have to look it up.
The data for the MAP is confusing, ambiguous, and there's a range of numbers, then there's sets of different ranges, and then there's a RIT score, that has a range, and there's another score that doesn't have a range...No, it does not help me at all.

NWEA does provides a resource to go along with the MAP test, called DesCartes, which is supposed to translate the RIT scores into more “usable” information for the teachers in their lesson planning. According the NWEA documents, DesCartes is “aligned to state standards [and] is designed to help you translate the raw data from your students’ assessments in to actionable plans for instruction . . . DesCartes orders specific skills by achievement level” (Northwest Evaluation Association, 2010). This resource provides charts for interpreting the RIT scores, but the teachers have to reference these in a separate notebook as they simultaneously look through the score reports. DesCartes is designed to be the answer to the teacher complaint that MAP results do not align with state standards and, more specifically, is supposed to give specific, concrete areas for instruction based on RIT scores in the subcategories for the MAP test. For example, a student with an RIT score in the range of 161-170 in the subcategory of number sense should be able to identify graphic representations of basic fractions.

Although presented by NWEA as providing these more specific data, the teachers need more than the general types of questions in the four subcategories to identify student needs within the numerous state standards. Because this resource is not really item or standard-specific

and because, for every student in each administration of the test, the teacher would need to look up the types of questions for the students' scores in all four subsections, this resource is rarely, if ever, used by the teachers. In fact, only one teacher mentioned DesCartes by name in the interviews and two other teachers made a reference that this resource exists. Similarly, the two instructional coaches did refer to DesCartes, although only one did so by name. These comments are presented in Table 4.9.

Table 4.9

Teacher and Administrator Comments about DesCartes

Teacher Comments	I know I don't use DesCartes the way I could. It's just so cumbersome, I think – so much information, I just feel overwhelmed. But I think if we really had more time . . . It gives us so much information to really work with it, I think it would be useful, more useful to me than . . . just getting that snapshot that we kind of rely on a lot of the time.
	I think there's some kind of page after page after page that you can go and find it, but the regular teacher doesn't have access.
	I can only go by what I've got in my handbook that I flip through like really quickly. That's all I can tell and . . . what I would show to a parent [if] they wanted to know what type of questions were on there. My kid missed this type of question, what does that look like? Then I show it to a parent, that's about it.
Administrator Comments	They can look at the Descartes and see what the questions are kind of like . . . It gives a lot of information, but it's, again, a lot of information. And so, it's very tedious to have to go through that for every single child under every single strand.
	There's a notebook over there about . . . how if a kid is scoring at this range . . . in this area, what is the best next step instructionally? We had not focused on that.

In all five instances, the comments included remarks on the inaccessibility or lack of helpfulness of DesCartes, especially due to issues of time, limited access, or the burdensome nature of using it. One teacher mentioned not having access, but another teacher showed me the binder on her bookshelf, and a copy is available in the teacher planning room. Whether or not there is a true lack of physical access, there seems to be a clear lack of usability to the point where neither the administrators nor the teachers use DesCartes or, in most cases, even mention that it exists.

The teachers do spend time, however, analyzing the data together in grade-level meetings with an instructional coach. As discussed in Chapter Two, the data from the state tests (the KCAs and practice KCAs) are used more often, but the MAP results are included in these discussions. Excerpts from the teacher interviews that explain what they do with the MAP data they receive are presented in Table 4.10.

Table 4.10

Teachers' Comments on Analyzing and Interpreting MAP Data

We just use that based on improvement, or non-improvement . . . We don't really go into why we don't think students are improving, or . . . what we could do to improve or help them improve. We don't talk about that.
Usually we discuss whether students need to move from flexible group[ing] and if we need to change what level they're at. We might discuss what was going on that day, like if a kid has a drastic drop what could have been going on that could have caused that kid to drop like that?
What we do is we take the MAP scores and we look and see where our students are and then we're looking at what kind of gains they're making. Then we, as a team, we look at our pacing.
[T]hey make me do data analyses of it, and then it goes in the notebook to keep them happy . . . We look at where the kids are low, but, you know, when you look at where they're low on the MAP, it's where they are on the state assessment so we just tend to go towards whatever the KCA has given us . . . We look for growth and set goals.
We just try to show them progress. We want them at least to progress to the point level that they give us. They give us certain point levels that they need to achieve by the end of the year.
We discuss it, we discuss it. . . Whatever [the instructional coach] asks us . . . Our strengths, our weaknesses, and you know, pretty much that . . . What we need to work on next year.
I think there's more emphasis on scores. Everything is scores, scores, scores, scores. It's not, "Has the child really learned it?" but "Can the child reproduce it on the test?"
We get together and we discuss it more, but I wouldn't say we agree on something. But everybody does their own thing. Nobody really follows—I mean, if somebody follows through with what was on the plan, there are three other people that aren't. So, it's face value. It's to make our [administrators] get out of our face.

Many of the teachers complained that they do a cursory analysis of the MAP data simply because they are told that they have to do this, but they are not truly using this information in a direct way to inform their instruction. One instructional coach explained in her interview that, in the younger grade levels where there are not yearly state assessments, they do use the MAP results more "because we don't have another standardized source of data," but for grades three, four and

five “we more tend to use the state standardized test data.” When MAP data are discussed, the teachers often do not see the value in it because they see the direct applicability of the KCA data to their instruction while the MAP data’s relevance seems more obscure. In general, the teachers look to the MAP results to show growth, but the KCAs show them what to focus on in their instruction; therefore, MAP data are quickly analyzed and put aside while KCA data are continually analyzed and directly applied to classroom decisions.

Qualitative Results: Observations

Observations during the MAP testing supported this focus on MAP scores and a desire to see growth. These observations also revealed more areas of confusion for understanding MAP scores. For example, during testing, one third grade teacher asked me about how to interpret the RIT growth charts and what students at that grade level should expect to see in gains. This teacher was also frustrated with the growth charts, questioning whether the chart was showing expected gains from the Fall administration to the Spring or just from the Winter to the Spring. Even more generally, this teacher said that she was “confused by MAP testing” and asked where she could learn more about it. Similarly, a fourth-grade teacher was talking with me about being given a single target RIT score (of 211) for all students at her grade-level as well as the same overall goal for growth (17 RIT points); she expressed that she was unsure that these were reasonable expectations for all of her students.

This desire to see growth was most clearly demonstrated by the post-it notes with goal scores that students would bring with them into the computer lab for MAP testing. As discussed earlier in the chapter, four of the teachers provided the students with these post-it notes and scores; the goal scores (at all grade levels for all students) were consistently five points higher than their previous MAP scores. One teacher specifically told me, “But what I do is take their

last score and add five points to it and give it to them on a sticky note and they stick it to their computer to see if they can meet that goal.” Even the teachers who did not send their students with post-it notes made mention of these goals to their students reminding them to think back to their last test score and to try to improve on it; it was clear in several classes that, even if the students did not have scores with them in the computer lab, they had been told what score they were supposed to try to “beat” from their previous testing. Many classes were asked to write down their scores as they finished to report back to their teacher or to record for themselves back in the classroom. Some of the teachers also put a line on the post-it notes where students were to record their new score as they completed the test.

Other teachers wrote down the student scores on a roster as the students completed the last question and their score popped up on the computer screen. As I passed by one teacher in the hallway whose class had just finished testing, I said hello and she immediately showed me her roster so that I could see that ten of her fifteen students that had tested that day had improved. Many teachers gave me similar, general updates saying that their class did really well or that not many met their goal scores. In many ways, this was the heart of their analysis – was there improvement and did their students make the five point gains they had set as their goal? Neither of these questions, however, address how to improve instruction or even correspond with the ways that the NWEA presents that the MAP data should be used.

Analysis

The teachers demonstrated less accuracy in their prediction of the student scores than was expected. However, overall, the teachers’ predictions were reasonable to their understanding of the MAP test. Their lack of accuracy in prediction is either due to the fact that they do not know their students well enough to predict their scores or it is due to a lack of understanding of the

MAP test and RIT scores. The data cannot decisively say which factor better explains the teachers' inaccuracy in predictions. However, based on the pattern of underestimating student scores, the RIT true and false survey results, and the qualitative factors that emerged when examining the individual teacher differences in predictions, the most probable conclusion is that the teachers are misunderstanding the MAP test and RIT scores.

Underestimation of Student Scores

The first piece of evidence that supports this conclusion is the consistent underestimation of the students' scores. The underestimation of growth is an interesting finding that stands out in this study. In particular, the increased error among third-grade teachers deserves attention. While experience with testing and assessment data can influence this result, the data provided point to the misunderstanding of RIT scores as a critical underlying factor to this error. As seen in Figure 4.2, teachers did not understand that target growth in RIT scores differs by individual students based on their current grade-level and previous scores. According to NWEA-MAP target growth charts, third-graders are expected to achieve greater gains in their RIT scores during the year than students in fourth or fifth-grades. However, since the teachers do not know to expect these larger gains, then it makes sense that their average predicted growth was lower, matching that of the upper grade levels.

The teachers' lack of familiarity with the target growth charts also helps to explain the overall pattern of underestimating student growth. As seen through the testing observations, teachers often gave their students a goal score going in to their MAP testing that was just five points higher than their previous score. However, as previously explained in the results section, this does not match how NWEA recommends that teachers calculate target growth and is less than any of the target growth charts suggest for these grade levels.

RIT True and False Survey Results

The area of greatest concern within this chapters' results is the extent to which the teachers do not understand RIT scores and, therefore, cannot be gaining much useful information for instruction from these misunderstood data. In fact, it is fortunate that the teachers report not using these data to inform their instruction since their misunderstanding of the data could lead them to make instructional decisions that are even harmful to the academic growth of their students. This is especially true since the teachers do not realize how much they misunderstand about the MAP data. Many teachers reported on the survey that they feel prepared to apply the results to their teaching, yet the RIT true and false items from the survey, the teacher interviews, and the observational data all show that the teachers are not adequately prepared to identify appropriate levels for student growth. The survey also showed that the teachers do not understand the ways that MAP acquires and reports student achievement levels. Additionally, the teachers trust the MAP scores too much; they have much greater confidence in the reliability of the scores than NWEA claims these scores to hold.

Individual Teacher Differences

This chapter's results on individual teacher differences provide a third area of evidence demonstrating that the lack of accuracy in prediction is more likely due to teachers' lack of understanding of MAP testing than due to a lack of familiarity with their students' mathematical achievement levels. The individual teacher differences in the ability to predict student growth on the MAP is particularly interesting in that many of the factors (years of experience, teacher efficacy, outcomes expectancies, student engagement, and cognitive activities in lessons) that one might assume would distinguish "good predictors" from "poor predictors" did not emerge as relevant in this study. The four teachers who were more accurate in their predictions did all

score above the average teacher score (of 48 percent) on the RIT quiz in the survey. They are, therefore, all somewhat knowledgeable about the test itself, but their understanding of the RIT scores was not the distinguishing factor between these teachers and the teachers who were less accurate in their predictions. Instead, two somewhat surprising factors emerged from the qualitative data to distinguish these four teachers from their colleagues: the use of post-it notes given to students with their goal scores during MAP testing and a general attitude of frustration with standardized testing. These teachers are most likely more aware of the MAP data than their colleagues since they are definitely interacting with these data enough to be looking up the students' previous scores, writing these down, setting goals, and discussing these data with their students. This greater awareness may make them better at predicting future scores because they are more familiar with MAP data, or their awareness may also make them more conscious of their students' patterns of achievement, in general, which then allows them to make better score predictions. These teachers' frustrations with standardized testing may also be, in part, due to their greater interaction with the data. Two of these teachers mentioned in their interviews that there is "too much" data, (a comment shared by many of the teachers in the study as described in Chapter Three). Similarly, these teachers may just be more frustrated by standardized testing because they are already more familiar with their students' achievement levels in mathematics and, therefore, are receiving information back from the tests that they do not need.

This finding might seem counterintuitive. However, it is the teachers who are most informed about the test results through their interaction with the student data that are most able to see the futility of their efforts. In other words, this is an example of the cliché, "familiarity breeds contempt." The teachers who are less able to predict student outcomes are more likely to feel that they are missing some piece of information about their students that a standardized test

might be able to provide to them. The teachers who are more familiar with their students' results on the MAP, who interact more with the data, and who can, therefore, more accurately predict the scores, then have more confidence to be able to evaluate the lack of applicability of the results to their students' instructional needs. Therefore, they are more frustrated with the use of standardized tests because they are in a position to better evaluate the lack of help these tests actually provide to them and to their students.

New Information the MAP Test Could Provide

Despite the low predictions, the teachers did do fairly well in predicting their students' scores. The teachers' predictions of their students' performance in the subcategories – predictions that they made without any reference to previous performance of the students in these subcategories – point to the specific areas of mathematics where the teachers best know their students' abilities. The teachers seem to be most familiar with their students' achievement in data and algebra. This suggests that the MAP assessment is not providing teachers with much new information about their students' strengths and weaknesses in these areas.

However, their less accurate predictions in geometry and numbers and computations may point to these areas as those where teachers do need more feedback on their students' abilities. Therefore, it would seem that the MAP testing provides some needed information to the teachers in these particular areas. While the test gives needed attention to these broad areas for the teachers, the specific areas of need within these categories is much harder for the teachers to target with MAP testing, as the test does not report specific standards needing attention or types of questions missed (as other assessments used by these teachers often do). In their interviews, very few of the teachers mentioned DesCartes – a tool provided by NWEA to provide this more specific information to the teachers. However, the teachers that discussed DesCartes as an

option for getting this needed information also explained that they have not used this resource much, if at all, and do not have the time to look up and sort through that information.

Overall, the results of this chapter show that MAP has the potential to provide teachers with some new information about their student mathematics achievement levels. However, that potential is not being realized for several reasons. The most important finding is that the teachers do not understand MAP testing, RIT scores, and, in particular, expected growth in RIT scores. The teachers know that they lack familiarity with the test, but they are not aware of how much misunderstanding they bring to the data they receive from the test. Without understanding the data, MAP testing cannot provide useful information to the teachers. Furthermore, the teachers report little use of the data, looking generally only for gains and using other standardized assessments to direct their instruction. With assessments in place that more directly match the instructional goals of the teachers, they are not motivated (nor do they have the time) to dig through the complex MAP data and related resources, such as DesCartes, to use these data in their planning. Instead, as seen in Chapter Three, the teachers express resentment as they are being required to discuss these data and are being held accountable to these data without much follow-through or relevance to their daily classroom teaching experiences.

Chapter Five

Education Policy and Additional, External Standardized Tests

This study looks at one test, in one school, in one school district, in one state. What can a glimpse at the use of the MAP testing by eleven teachers in this one context tell us about education policy that is often being determined at a national level? I believe that this case study provides more important, substantive information than a larger study of national policy that neglects the experience of the individual teachers and students who are affected by policy decisions at all levels. This chapter will, therefore, broaden the discussion from the context of this one school and this one test to the greater field of education policy and will discuss the implications that what is learned from this study can and should have on policy decisions.

Summary of Findings

As seen in this study, testing and accountability have infiltrated and defined the culture (the beliefs and behaviors) of the school. Current research speaks to the culture of testing that is present throughout schools in the United States. From this case study, the concerns that develop amid this culture of testing can be broken down into three main categories: the culture of data, the culture of pressure, and the culture of confusion.

The Culture of Data

Chapter Two reveals discrepancies in the intended purposes of MAP and the ways that MAP data are actually used in the school. Rather than serving as formative data to inform instruction during the year, MAP testing has become an interim summative assessment leading up to and even following the yearly state test. Since the Spring administration of MAP occurs after the state assessment, the teachers view it as both an unimportant, additional assessment (since the “big” state test is already done) and as an unfair indicator of teacher quality that

administrators will use to judge their year's effectiveness with their students. In particular, the teachers may be justified in seeing this as unfair, not only because placing so much weight on a single test is an unfair use of testing data, but also because the teachers' outcome expectancies were significantly lower than their mathematics teaching efficacy scores, indicating that even the teachers who believe they can teach mathematics well do not believe that they have much control over the students' performance on mathematics assessments. Chapter Two also shows how the culture of data present at this school produces a mindset that more data are always inherently good and necessary. The teachers, however, indicate that they have reached a saturation point with the data that they are receiving, especially since the teachers do not find the data acquired through MAP testing to be applicable to their instruction, generally, and even more so as they compare MAP data to data that they receive from other practice tests more closely aligned to the state assessment and state standards.

The Culture of Pressure

Chapter Three describes the culture of pressure that is present in the school. While no one in the school is immune from the pressure related to assessments, the teachers are the mediators of this pressure since they receive (and internalize) it from the administrators and then must filter how it is passed onto their students. MAP testing becomes one (or, actually, three) of the many instances that this pressure is heightened; the greatest pressure, of course, comes from the state's annual KCA testing. MAP testing interacts with KCA testing, both in the understanding of its ability to predict KCA outcomes and its assumed role as preparation (testing practice and instructional guidance) for KCA. More importantly, however, enough weight is placed on the MAP scores themselves as evaluative of the teachers' instructional quality that the MAP carries a great deal of pressure for them independent of the KCAs. This is true for the

students, as well, when they are aware of the ways that their MAP scores will determine educational placement decisions about them.

The Culture of Confusion

Chapter Four highlights a unique piece that MAP testing brings to the culture of testing – confusion. Teachers (and administrators) do not understand MAP testing and, especially, do not understand the data provided as MAP results. Some teachers recognize and admit their confusion with these data. Other teachers believe that they are competent in understanding these data, yet their survey results reveal deep misconceptions about MAP testing and misinterpretations of the meaning of RIT scores. This misunderstanding of RIT scores and of the expected growth over time with MAP testing seems to account for the teachers' consistent pattern of underestimating their predicted student scores. Those teachers who did predict their scores well (i.e. with no significant statistical difference between their predictions for student growth and the students' actual growth) were those teachers who likely interact more with the MAP data but who also express skepticism toward standardized testing in general. Adding to the levels of confusion are the resources provided by NWEA to clarify how to use MAP data; even the teachers and instructional coaches who were more aware of NWEA resources for helping interpret MAP scores and apply them to instruction find these resources to be inaccessible, overwhelming, or just too time-consuming to use.

Strengths and Limitations

This study provides an in-depth analysis of the effects of testing and accountability policies as they are carried out at the level of implementation – the classrooms. Specifically, this research elevates the experiences of the students and teachers as the area of greatest concern for truly understanding the effects of additional, external standardized tests. Previous research has

tended to look at large-scale effects of testing, generally, and has also focused only on the traditionally high-stakes tests, such as the state assessments. By looking at an example of additional, external standardized assessment whose use is mandated in the school only by a decision of the local school board, different policy dynamics are revealed. The MAP test is both a response to national policy and a mechanism of local policy; the implications of this dual role are messy and difficult to distinguish from the broader culture of testing and accountability. However, an attempt to dissect this culture of testing into its components at work in a particular school gives insights into how these types of policy decisions are (or are not) meeting their intended purposes, how they create a new set of issues to be addressed (or at least acknowledged), and how new directions in educational policy could better support the educational experience of teachers and enhance the educational benefit to students.

In this study, most of the data are obtained directly from the study's participants, those most affected by education policy and by the implementation of MAP testing, specifically. A particular strength of this research was the existing relationship that I had with the teachers and administrators at this school, giving me physical access to the classrooms and data but also personal access to the less inhibited thoughts and concerns of the teachers and administrators about the testing, the instruction, and the educational atmosphere at the school. Similarly, my prior experience in the school allowed me to more easily identify the school norms and recognize when something I was observing was exceptional to the regular flow of the day's interactions and activities.

A limitation of this study is that, being a case study, it examines a single case – one school, eleven teachers, and three administrators. The case study design does limit the generalizability of findings as understood in a traditional sense. For example, the quantitative

findings are mostly descriptive since the sample sizes were often too small for more in-depth statistical analysis. However, it is this “smallness” that allows for the qualitative pieces to bring the story of testing at this school to life. Therefore, the limited size of the study’s participants does not limit the generalized understanding that one can take from this study and apply to similar contexts. The application of these findings to other contexts will never be (and is not intended to be) an exact fit. Rather, this research provides a perspective through which future researchers can approach similar contexts and gain further insight into the culture of testing and, in particular, the cultures of data, pressure, and confusion that may be present within the culture of testing.

Recommendations

Discontinuing Use of the MAP

MAP is a well-constructed test that does a good job of measuring growth over time and identifying students’ mathematics achievement levels outside of the constraints of testing only grade-level standards. This is what NWEA claims that MAP is and can do. As with any test, one can recommend improvements or other ways that it can be more helpful, but the problem with the use of MAP testing is not truly a problem with MAP. While the MAP assessment may be a valuable test for many reasons, it is not beneficial when it is another, *additional* test in an environment that is already saturated with data and overwhelmed with testing.

The school selected for this case study is over-testing, with students completing computer-based standardized tests for mathematics weekly. Much of this testing – the practice KCAs, in particular – is provided by the state, recommended by the district, and then required by the school-level administration. (The MAP test is not a state-level decision; rather, it is both purchased and mandated at the district-level.) The teachers do not have any option as to whether

or not their students will, then, participate in all of this testing. The decisions are handed to them, and they have no choice but to comply with these decisions. Therefore, these policies, that are supposedly intended to help them as teachers, blatantly ignore the data's lack of utility to the teachers, the additional pressure on students and teachers that are a direct result of these tests, and the confusion caused by the results that are, in practice, often meaningless to their instructional decisions. The most important recommendation based on this study is that classroom-level decisions be restored to the classroom teacher so that the teacher can choose assessments that are meaningful to and useful for their instruction. Practice KCAs and MAP testing can be provided as *options* to the teachers, but should not be required; and all teachers should be cautioned to avoid a situation of over-testing, i.e. choosing to assess more without gaining any *useful* data for their instruction.

Too much testing is not helpful for anyone, even those administrators who think that they cannot have too much data. One has too much data when the data provide little or no new, useful information. At the beginning of this study, I thought that a potential downfall of MAP testing was that it was not providing *new* information to the teachers, and, therefore, I asked the teachers to predict student scores to see if they would already know the achievement levels of their students. This study provided mixed results on that particular hypothesis (for various reasons discussed already in this chapter and in Chapter Four). Instead, this study provided much evidence that the data from MAP testing are not *useful* to the teachers, whether or not these data are new sources of information for them. Therefore, it is the lack of utility of the data (or, at least, the lack of ease for utilizing the data) that becomes the strongest argument for discontinuing the use of the MAP in the specific context of this school.

That does not mean that the MAP test could not be useful for some purposes in some contexts. But, it does mean, that in similar contexts, where assessments are already integrated throughout the school year and the teachers are already using other data to inform their instruction, the potential benefits of MAP testing could rarely outweigh the negative effects of additional testing. This is especially true because the results of MAP are more difficult to interpret, both because of the use of obscure RIT scores and because the breakdown of the data into the four subcategories of mathematics is still not enough to isolate the specific needs of students, at least not without substantial efforts to decode the scores in each category for every student using the cumbersome DesCarte resource provided by the NWEA.

Considering Alternative Forms of Assessment

Before considering alternative forms of assessment, researchers must consider what they truly desires to assess. Many school districts desire to have multiple assessments in order to reduce the chance that a student's scores reflect a bad testing day or even a biased test; they see these multiple assessments as a way to triangulate data on student achievement and, therefore, obtain a fuller picture of the student. Three different standardized test scores, however, is not what is meant by "triangulation" of data; triangulation cannot be accomplished by looking at three versions of one form of assessment. Instead, the data themselves should be in three forms, such as a standardized assessment, a student portfolio, and observational assessments of students actively engaging in mathematics problem-solving.

As with many schools, the school in this study does not need another measure of student growth or mathematics achievement levels; therefore the MAP test is not an appropriate form of assessment. In reality, the school is using MAP testing primarily in two ways: (1) as a predictor of student achievement on the state assessment and, therefore, as a predictor of the school's

chances of meeting AYP and (2) as a measure of teacher quality. Once these purposes are honestly identified, then appropriate assessments can be considered.

The first main purpose, then, is predicting AYP. From an admittedly cynical standpoint, predicting a school's chances of meeting AYP is getting easier and easier; as schools move closer to 2014, when all schools are to have 100% proficiency in all subgroups, there is almost no chance of meeting AYP. Therefore, no assessment (other than the assessment of the impracticality of NCLB) is necessary. More reasonably, however, the school only needs to predict scores for the "bubble" students – those who were close to meeting proficiency or just barely met proficiency the year before. Another test is not necessary to identify these students; the school already has the data to know which students fall into this category. I am not advocating the current trend to teach specifically to these students as a strategy for making AYP, but I am stating that, if this is the case, subjecting all students to MAP testing in order to predict the scores of these "bubble" students is, at the very least, inefficient and, more importantly, potentially unethical.

This leaves the second often-ignored, but widely-acknowledged use of MAP testing at this school – judging teacher instructional effectiveness. While the administrators openly talk about using the MAP results for these purposes, I do not believe that they would choose MAP testing as the ideal measure of instructional effectiveness. Corresponding to the culture of data, the administrators want objective (generally, quantitative) information that they can use to judge quality. However, teaching effectiveness is a complex construct that cannot be measured by student test scores alone. In fact, in many cases, student test scores may greatly misrepresent the quality of teaching taking place in a classroom. More subjective measure of teacher quality would also be more consistently valid – frequent classroom observations, teaching portfolios,

and even peer teaching reviews – and better serve the purposes of judging the effectiveness of a teacher.

Questions for Future Research

The research presented in this study can be enhanced by further research in a variety of areas within the fields of assessment, mathematics education, and education policy. Most specifically, more research should be conducted apart from the NWEA’s own research on MAP testing and should further explore the experiences of students, teachers, and administrators using this test in other schools. Likewise, more research on additional, external standardized testing is necessary to see the commonalities between the experience of schools using MAP testing and those choosing other external assessments for similar purposes.

An interesting area for future research would be an exploration and identification of which students the teachers’ *can* predict growth well and of which students the teachers are poor predictors. With a larger sample, quantitative analysis could reveal subgroups of students within the population that the teachers know better. Are there characteristics of students about whom teachers need more data to better understand their instructional needs? For example, the correlations in Table 4.2 between teacher predictions and student achievement are strong, but also clearly demonstrate a lot of room for improvement in teacher prediction. If future research can identify the students who are being “missed” in these correlations, then assessment tools, such as MAP, can be used in a more intentional fashion to give teachers specific data that they *do* need. Similarly, with a larger sample of teachers, perhaps future research could aid in indentifying more generalizable characteristics of teachers who are better predictors.

Another way that this study, itself, could be improved upon in future research would be by having teachers predict student scores on both the MAP and the KCA tests. Most likely, the

teachers would do much better in predicting KCA scores, indicating their knowledge of student achievement levels *and* their familiarity with the KCAs. This would then allow for a more definitive discussion of how the confusion around MAP testing and RIT scores interferes with this assessment's utility for teachers. Operating under the well-supported, although not conclusive, assumption that RIT scores are not a teacher-friendly form of data, future research should also include studies of ways to improve teacher understanding, use, and application of these scores. Perhaps, more realistically, researchers could also suggest more meaningful measures for teachers that still retain the RIT scores' main strength of demonstrating growth over time.

More broadly, more case studies that focus on the culture of testing within schools would benefit policy makers in two ways. First of all, schools experiencing similar problems with too much testing, too much data, and too much pressure (like the school selected for this case study) are easy to find; obtaining consistent findings in similar contexts is important to demonstrating the pervasiveness of these problems. Secondly, schools that have a more unique response to the movement of accountability and testing can also provide important information on ways that administrators and teachers may be able to better respond to the pressures being placed upon them. However, this research must be chosen and reported with caution – these studies must focus on schools with similar contexts; they should not be anomalous because of their structure (charter schools, for example) nor because of their demographics (student backgrounds, district funding). The experiences of the teachers and students presented in this study should only be compared to those in other urban public schools who also have struggled each year to meet AYP, who have a high turn-over of students within and between districts, and who serve a large population of English language learners. With these conditions in mind, examples of schools

that are operating outside of the culture of testing would contribute greatly to a fuller understanding of how education policy is affecting the educational experience at the level of implementation. How have these schools avoided the movement toward over-testing? How do they handle the pressures from outside the school for more data? How do they define accountability for themselves? How do they (or, simply, do they) measure teacher effectiveness? In what ways are they utilizing assessments, in all forms, to monitor and enhance student learning?

Shifting the Paradigm of Education Policy

While more questions regarding the purposes and effects of additional, external standardized assessments can and should be explored in future research, education policy concerns cannot be addressed by research alone. As long as policy-makers and researchers continue to operate within the current paradigm of educational thought, the same arguments will continue to dominate the discussion – top-down versus bottom-up reform; centralization versus decentralization of decision-making; traditionalist versus constructivist pedagogy. Instead, we must examine the current patterns of thought and reconsider the underlying assumptions of these discussions. We must find a way to take action in education in a manner that does not favor one extreme nor attempt to balance these opposing views but, instead, operates outside of these concepts that we tend to assume are mutually exclusive. This discussion on what a true shift in the paradigm in education policy could be will start with the particular piece of the paradigm related to testing and move to a broader perspective on the structures of power in schools.

Emphasis on Test Scores

The current iteration of the educational policy paradigm is characterized by test scores. Tests scores compare education quality and identify success at all levels – (un)successful

students, (un)successful teachers, (un)successful schools, (un)successful districts, (un)successful states, and (un)successful countries. Test scores, however, do not just serve the role of identifying success; they also define what it means to be successful. The fact that test scores are both the definition *and* the indicator of success is highly problematic. Nichols and Berliner (2007) discuss another similar, problematic dichotomy in the educational culture of testing: “The tests are seen by some as the perfect policy mechanism because they are both *effectors* and *detectors* – they are intended to *effect* or cause change in the system and then *detect* whether changes in the system actually occur” (p. 6). Therefore, not only do test scores simultaneously define and identify success, they also are supposed to be an agent for promoting success in education. Clearly, there is too much weight placed on what testing can accomplish. But, beyond that, this over-reliance on testing as the “cure all” within the realm of education policy has the effect of both corrupting the testing itself and the system as a whole.

Nichols and Berliner (2007) and Koretz (2008) both cite the inevitability of this result based on the principle of Campbell’s Law. Campbell (1976) states that “the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it was intended to monitor” (p. 49). The idea is that, by placing too much emphasis on the power of one indicator (such as a test score), the indicator will be corrupted to the point that it no longer measures what it intends to measure. In other words the indicator, (in this case, the test score) loses its validity (Koretz, 2008; Nichols & Berliner, 2007; Ravitch, 2010). In practice, this can include blatant corruption such as cheating or misrepresenting statistics, but it can also include less blatant corruption such as teaching to the test and narrowing the curriculum along with

related side-effects, such as eroding teacher morale (Nichols and Berliner, 2007). Nichols and Berliner explain (2007):

As Campbell's law predicts, the more importance that an indicator takes on, the more likely *it*, and the people who depend on it, will be corrupted . . . Under these conditions we must worry that the process that is being monitored by these test scores – quality of education – is corrupted and distorted simultaneously, rendering the indicator itself less valid, perhaps meaningless (p. 30).

What does an improved MAP score indicate? Even if the score has not been corrupted by one's ability to teach to the test (since the teachers indicated that it is difficult to teach to the MAP test), the teachers do not understand what the improved score is showing. The students, teachers, and administrators are all looking to these scores to see improvement, i.e. to see the number increase. If it does not increase, there may be more effort to understand why it did not increase, but there is little effort to understand the meaning of the score more deeply than whether it decreases, stays the same, or increases. The indicator, i.e., the score, is all that is important and, therefore, it loses (or, perhaps for these educators, never even had) meaning.

How MAP becomes "High Stakes"

One may argue that MAP is not a high-stake test. When viewed by itself, apart from its context of why and how it is used in the schools, this is true. MAP is not inherently high-stakes. However, MAP was adopted as part of the culture of high-states testing – to identify, predict, and prepare for the state assessment. Popham (2001) explains how additional standardized tests become high stakes:

. . . even if there are no tangible rewards or penalties . . . those tests still qualify as high-stakes measurements. Almost everyone wants to succeed, and teachers want their schools' students to score well. Few folks want to be at the bottom of what is seen to be a quality-ranking continuum. That's why there are few large-scale assessment programs these days that are not fundamentally high-stakes in nature (p. 34).

The pressures experienced by the students, teachers, and administrators in relation to this test indicate that the stakes are high, even when those stakes are less easily identifiable than they are in the tests required under NCLB. Interestingly, when teachers complain (or praise) that they cannot teach to the MAP test, they are commenting on the fact that these test scores are less “corruptible.” However, as part of the culture of high-stakes testing, it is still subject to abuse and misuse. While not as extreme as other testing cases, MAP testing is subject to Campbell’s Law in that too much emphasis is placed on the indicator as the goal itself, rather than maintaining a goal that the indicator helps to inform.

The consequences of that are seen in this study. Teachers expressed concerns about teaching to the test and about the narrowing of the curriculum to meet what is on the test. These concerns apply to the broader testing culture at the school, not just the MAP test. In fact, several teachers pointed out that it is difficult to teach to the MAP test and that it covers material that the state test does not assess. One might, conclude, therefore, that the MAP test helps to alleviate these two concerns. However, this would be an incorrect interpretation. Additional testing that does not inherently restrict the curriculum or allow students to be directly prepared for each type of question does not negate these effects that are already present from the state test. Instead, the MAP test actually reinforces some of these negative effects from the true high stakes tests that are present. For example, MAP assessments are given in mathematics and reading; these are the two subject areas in which NCLB mandates annual testing. A narrowed curriculum that increases time and attention placed on mathematics and reading also decreases time and attention placed on other subjects, such as social studies. MAP testing reinforces that only mathematics and reading are the important measurable outcomes of the education happening in the classroom and, therefore, helps to solidify this curricular narrowing. Rothstein, Jacobsen, and Wilder

(2008) comment extensively on this narrowing of the curriculum due to the testing emphasis on mathematics and reading, stating that “holding schools accountable for math and reading test has created incentives for educators to pay less attention to curricular areas for which they are not held accountable” (p. 9). Even the practical aspects of administering MAP reinforce this narrowing at the school presented in this case study since, for most of the classes, the students’ regular time for technology education was replaced by testing on the computers.

Similarly, MAP results, just like the results of the true high stakes tests, are used to make evaluative decisions that affect students and teachers, and the pressure that is associated with these results is palpable at the school. If the stakes were not high, a student would not burst into tears when she sees her score; for that student, who was looking to qualify for the gifted program, the stakes are high. Certainly, this is not as “high stakes” as being promoted to the next grade, but it is still a significant decision that the student knows would be decided based on her score. The teachers also feel immense pressure from MAP results. While MAP does not determine AYP, so it again is *less* high-stakes, the teachers know that they are being judged on these scores to the exclusion of other evidence of their teaching competence. When student and teacher morale hang in the balance of the outcomes of these scores, the stakes are already too high.

Caring About the Individuals, More than the Big Picture

Were the culture of high-stakes testing not already part of the context of the school, the administrators might be less likely to place so much weight on the results, the teachers might be less likely to feel so judged by the scores, and the students might be less likely to realize how this ninety-minute activity affects their educational trajectories so much more than the countless hours in the classroom that they spend learning. However, when administrators and teachers

know that AYP has immense consequences and they see MAP testing as a way to predict AYP, then MAP scores are given much more power than they are intended to have. Consequently, students are given the burden of producing an anticipated score rather than the responsibility of learning the important content for their lives outside of testing. This is how too much weight on a test score corrupts the environment and the individuals operating within that environment. While Campbell's law does predict a validity problem, it also predicts a corruption of the social processes themselves – in this case, a departure from caring about the lives and education of individual students in individual classrooms in order to appease public and political interest in the test scores. McNeil (2000) agrees that “standardization, as it becomes institutionalized” results in “the decline in the quality of teaching and learning” and further emphasizes that these consequences “are not aberrations or malfunctions in the system but *the logical consequences of the system when it is working*” (p. 231).

Porter and Chester (2002) suggest that symmetric accountability systems provide a means for responsibility to be shared among all those involved in education (students, staff, and policy makers); specifically they define “symmetric accountability” as programs that provide balanced incentives for students and for schools to demonstrate gains. Linn (2003) also upholds the need to switch to a system of shared responsibility for improvement to truly occur in education. Similarly, Noddings (2007) contrasts the concept of *responsibility* with that of *accountability*, explaining:

Responsibility is a much deeper, wider ranging concept than accountability. Typically, a worker or teacher is accountable to some higher authority, and accountability can often be satisfied by conformity, compliance with the letter of the law. In contrast, responsibility points downward in the hierarchy. As teachers, we are responsible for those below us – those for whom we serve as authorities. Teachers may be *accountable* to administrators for certain outcomes, but they are *responsible* to their students for a host of outcomes. Many of these outcomes are not easily measured (p. 39).

One of the problems with current education policy is that it is designed to target groups rather than support individual learners. No Child Left Behind actually calls for no *subgroup* (rather than individual child) to be left behind, i.e. accountability of test scores is broken down to identify specific subgroups (generally, minority groups) of students and their achievement. Of course, there is value in bringing to light the inequities within the educational system with regard to under-represented and previously ignored subgroups of students; there are also dangers in disaggregating the data in this way so that certain subgroups are “blamed” for keeping a school from meeting AYP.

Beyond these often-discussed pros and cons, however, is the underlying problem of continuing to ignore the individual. Education happens in a group setting, but it occurs at the level of the individual. Fundamentally, education rests on the interaction of individual students with individual teachers in an environment of responsibility. That does not have to exclude the concept of accountability; in fact, it should not exclude such an important concept since those personal interactions of teachers and students do not occur in an isolated bubble but rather in the context of a system that cannot be ignored. However, when bureaucracy becomes ingrained into a school and its classrooms and the need to nurture individual children is pitted against the need to meet public expectations and political requirements, then that school “is thus structured to be in conflict with itself” (McNeil, 2000, p. 11). Nichols and Berliner (2007) explain the problem with the current accountability structures this way: “When rigidity replaces flexibility and cruelty replaces compassion, we see our educational system in decline. This decline is not due to the promotion of accountability but the promotion of a faulty accountability system” (p. 78). When the system requires the focus to be weighted more toward accountability than responsibility, then the system has failed the individuals it is designed to serve.

Similarly, Elmore (2004) explains the value of internal accountability systems over external policies and external accountability systems, specifically in terms of the ways that internal accountability systems engage the concepts of individual responsibility and collective expectations. He explains that the internal system has a greater influence on behavior because it “corresponds closely with teachers’ understanding of their personal responsibility” in the classroom (p. 175). In other words, the teachers’ understanding of their personal, individual roles as agents in the classroom inherently brings with it a sense of responsibility that can work in a complementary fashion with internal accountability systems and expectations at the school-level. Elmore further explains, however, that when external policies conflict with this internal system, the internal system is the stronger influence, with teachers protecting their personal sense of responsibility. Furthermore, Elmore suggests that education policy should be adopted with this understanding of internal accountability so that policy makers do not waste efforts attempting to impose policies that ignore these systems, explaining, “the effects of policies are determined . . . by the ways individuals and organizations respond to them . . . [S]ystem-level policy makers and administrators should base their decisions on a clear understanding of the results they want to achieve in the smallest unit – the classroom, the school” (p. 4-5).

Organization of Schools and Subsidiarity

As with all education reform, this is not a new issue, but it is presented in a new context (Tyack & Cuban, 1995). Rather than be caught in another pendulum swing of education policy, we need to create a new paradigm through which we view education policy. This is not a matter of being more or less traditionally-minded in education. As Noddings (2005) explains, “. . . [T]his heavy emphasis on testing should not be confused with traditional or classical education. There are many thoughtful, traditional educators who share my disgust with the current

trivialization of education” (p. xiii). Nor is this a matter of being more focused on bottom-up or top-down reform. Similarly, it is not a matter of centralization versus decentralization of decision-making. Schumacher (1973) speaks about the pendulum swings in large organizations, alternating between phases of decentralizing and centralizing, and he suggests, “Whenever one encounters such *opposites*, each of them with persuasive arguments in its favour, it is worth looking into the depth of the problem for something more than compromise, more than a half-and-half solution. Maybe what we really need is not *either-or* but *the-one-and the-other-at-the-same-time*” (p. 259). Therefore, I am not calling for an extreme or even a balance between the extremes of reform; instead, I am calling for a paradigm shift in the structure of education policy and administration based on the principle of subsidiarity.

Subsidiarity is a social principle that has been most often used in political discussions in Europe but has been applied to economics, organizational theory, and education. Beare (1995) explains that the principle of subsidiarity means that “any collectivity, before it usurps the power vested in the local body, must show cause why it can discharge that function better, more efficiently, more humanely, more skillfully” (p. 147). Shumacher expands on the principle by explaining,

the higher level must not absorb the functions of the lower one, on the assumption that, being higher, it will automatically be wiser and fulfil [sic] them more efficiently . . . The principle of Subsidiary Function implies that the burden of proof lies always on those who want to deprive a lower level of its function, and thereby of its freedom and responsibility in that respect; *they* have to prove that the lower level is incapable of fulfilling this function satisfactorily and that the higher level can actually do much better (p. 260-1).

Messner (1949) addressed this principle in his writings on social ethics; since then, it has been applied to economics and organizations broadly (Shumacher, 1973) and education specifically (Beare, 1995; Coons & Sugarman, 1978; Macpherson, 1998). While somewhat obscure, it

answers the call for a response that is not another pendulum swing nor another cycling back to the previous uses of the language of reform. Subsidiarity can exist within complex, hierarchical systems such as the system of education. However, it establishes a respect for individuals and lower levels within the system and a preference for them to have responsibility over the processes that are most within their realm of control. Anything that they cannot handle well at that level can then be addressed in a complementary fashion through the higher levels of the hierarchy, with preference always given to the most localized level for handling the specific function. Messner's (1949) definition of this principle is "as much liberty as possible, as much [organizational] interference as necessary . . . The ideal . . . lies in the greatest possible measure of freedom for the individual within the framework" (p. 198).

Subsidiarity answers the educators who see the movement of educational policy as more centralized and nationalized, such as Conley (2003) who states, "[T]he state and federal role in education policy has become increasingly activist, and, when viewed from the local perspective, intrusive" (p. 1). It also maintains a respect for the individuals, acknowledging that "any attainment of the common good is ultimately owing to the activity of individuals" (Messner, 1949, p. 197) and "at the base of all social systems are relationships between people" (Macpherson, 1996, p. 8). But it does so while maintaining the need for a hierarchy and structure and even some "intrusion" of policy and mandates when the principle of subsidiarity is met. This aligns with some of the conclusions of Ravitch (2010), who explains:

[O]ur schools will not improve if elected officials intrude into pedagogical territory and make decisions that properly should be made by professional educators. Congress and state legislatures should not tell teachers how to teach. . . Pedagogy – that is, how to teach – is rightly the professional domain of individual teachers. Curriculum – that is, what to teach – should be determined by professional educators and scholars . . . acting with the authority vested in them by schools, districts, or states (p. 226).

Applying the principle of subsidiarity to the organization and governance of public schools would alleviate concerns such as that of Cochran-Smith and Lytle (2006), who call for education policy that reaffirms the “knowledge and understandings teachers derive from their local experiences and relationships with students” (p. 689).

For example, within the current movement for accountability and assessment, local control has been continuously stripped from schools (and even from states) in order to meet federal mandates. Elmore (2004) explains, “Prior to NCLB, ensuring accountability in schools had been primarily a state and local responsibility, with the federal government . . . playing a supportive role . . . Post-NCLB, the federal government has essentially preempted the field of education accountability” (p. 201). The question is whether the federal government met the principle of subsidiarity in mandating the accountability system of NCLB. Can the federal government prove that the lower levels within the education structure cannot adequately monitor educational quality *and* simultaneously prove that it can do the task much better? Do NCLB and related school accountability measures show that the federal and state governments’ policies are better, more efficient, and more humane? Many researchers questioning NCLB provide evidence to question this intrusion of the national government into the local-level school decisions. Even state governments, mostly in response to NCLB, are taking on roles in education that previously belonged to local communities. According to the principle of subsidiarity, this cannot be done solely based on an assumption that more centralized control results in more efficiency and better results. Instead, the humanity of the individuals that the system is intended to serve must always be preserved, and power should only be consolidated further up in the structure when there is evidence that the individuals can be better served by this

intrusion of power. As Nichols and Berliner (2007) comment, “It appears to us that the most important problem . . . is the loss of humanity that high-stakes testing produces” (p. 77).

Prior to NCLB, schools did have standardized tests and accountability measures in place, and teachers did use assessments to monitor their students’ progress and make informed instructional decisions. While there was room for improvement, this structure allowed for the tests to inform instruction. Now, as seen in this study, one test is used to inform the preparation for the next test, or, even worse, it is just another test to endure and it is virtually ignored beyond that. This case study, in particular, provides evidence that local school boards are overstepping their appropriate involvement in the daily functions of schools and classrooms, as MAP testing is a local district decision; this decision, then, is binding on each school and, consequently, on each teacher. The teachers no longer have the control to decide if MAP testing is the best tool to inform their instruction, even though that is stated to be its primary purpose. By taking away this decision and power from the teachers (where the decision for how to inform their instruction should professionally lie), the district has violated the principle of subsidiarity. According to this principle, the district could continue to purchase and provide MAP testing as an *option* to teachers who desire to use this assessment, but the teachers should not be *required* to use an assessment whose primary purpose is to inform their instruction. Following subsidiarity, the district should only require assessments (or other mandates on the students and teachers) that are necessary for appropriate district-level decisions.

Of course, the overuse of external, standardized assessment is in response to state policies and federal policies that also violate the principle of subsidiarity. Educational researchers critiquing the culture of testing note that “the need to test has replaced the need to care, a corruption of the traditional role of teachers . . . What we have observed are cases of apparent

injustice that occur when laws are written . . . a long distance from the places where the laws must be administered” (Nichols & Berliner, 2007, p. 73). The more that the principle of subsidiarity is violated, the more others feel the need to also violate it. Nichols and Berliner (2007) found this pattern for test preparation: “If one district engages in extensive test preparation, then all the districts feel the need to do so. Each district worries that they will look bad and be shamed in the press . . . So the rush is on the corrupt the indicator” (p. 124-5). As of 2006, MAP testing had been adopted by over 150 public school districts in Kansas; beyond that many private schools, individual public schools, and other educational entities within the state were also implementing MAP testing (Northwest Evaluation Association, 2006a). MAP testing appears to be part of the “rush” to “corrupt the indicator” as the decisions for monitoring student progress and the information collected for informing instruction are taken away from the teachers and put in the hands of school district administrators, those more removed from the instruction and from the individuals providing and receiving the instruction.

Concluding Thoughts

I am beginning and ending my dissertation writing in my own voice. In part, this is a conscious choice to frame my work as my own. All research is a result of efforts to work through an issue that is both personal and remote. The topic is personal in that there is a desire to study it, to bring it to the attention of others, to find answers worth finding. But the execution is remote – hidden in academic language in the form of an academic study that, in most cases, few outside of academia (or, for that matter, within it) will notice. Despite feeling disconnected from my research through much of the process of writing up this study, I have come to the point of realizing that this is a topic about which I care deeply, so I must introduce and conclude this work explaining why I care and why, perhaps, those who do read it should care, too.

I have no authority and little power to affect change in education policy. I do not believe the cause is hopeless; instead, I believe that no one person has the power to shift the paradigm of education policy as I am suggesting. Just as I approach education policy from the perspective that what matters most is that which happens with the individuals at the level of implementation, I also believe that it is only through individual conversions that the system, as a whole, can change. That does not mean that the responsibility rests on the teachers alone. Some educators have the responsibility to care about the individual students and teachers through making and revising national and state policies. I, however, am drawn to the individual students and teachers themselves. I came to study this topic and education policy, in general, so that I could help create a context for teaching and learning where the individuals had the freedom to care about each other and uphold each others' dignity within the classroom.

Educators and policy makers are all trying to make the best decisions as they see their choices presented. But we need to restore to teachers (and even to students) their right to authority in their classrooms and authority over their educational experiences in the many circumstances where they are the ones who have both the most relevant information to make the decisions and the most investment in the outcomes of the decisions being made. I do not necessarily trust our educational system, but I do trust our educators. I do not trust that they are all wonderful teachers or even that they are all wonderful people (although most of them are), but I trust that their desire to do good for those placed in their care is more powerful than my desire to do good from writing a dissertation. As I said as I began these concluding remarks, I have no authority and little power to affect change. Teachers have the greatest power to change education, and they should have much greater authority to do so than they are currently allowed.

I care about teachers and students and, in particular, I care about those I know – the ones with whom I have studied and taught and the ones whom I have studied and taught. I see the current culture of testing as an injustice against them. And so, for their sakes, I present this research and my thoughts. I do not believe that I have the power to change the current paradigm of education through conducting and sharing this study, but that does not mean that I believe any less in the importance of the change. I know that I, along with many other educators and researchers, care deeply about these issues. I hope that we can come to care even more deeply about how our responses to these issues can and do affect each other, especially those in the schools.

References

- Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice*, 42(1), 18-29.
- Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258-267.
- Baker, F. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Retrieved on July 31, 2010, from <http://echo.edres.org:8080/irt/baker/>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.
- Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education*, 51(5), 384-397.
- Beare, H. (1995). New patterns for managing schools and school systems. In C.W. Evers & J.D. Chapman (Eds.), *Educational administration: An Australian perspective* (p. 132-152). Melbourne: Allen and Unwin.
- Black, P., & Wiliam, D. (1998). Inside the back box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148.
- Blink, R. J. (2007). *Data-driven instructional leadership*. Larchmont, NY: Eye on Education.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Earlbaum Associates.
- Campbell, D. T. (1976). *Assessing the impact of planned social change*. Paper presented for the Public Affairs Center, Dartmouth College. Retrieved on July 19, 2010 from http://portals.wdi.wur.nl/files/docs/ppme/Assessing_impact_of_planned_social_change1.pdf

- Cochran-Smith, M., & Lytle, S. L. (2006). Troubling images of teaching in No Child Left Behind. *Harvard Educational Review*, 76(4), 668-726.
- Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology*, 78(2), 141-146.
- Conley, D.T. (2003). *Who governs our schools?: Changing roles and responsibilities*. New York: Teachers College Press.
- Coons, J. E., & Sugarman, S. D. (1978). *Education by choice: The case for family control*. Troy, NY: Educator's International Press, Inc.
- Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications, Inc.
- Darling-Hammond, L. (2003, February 16). Standards and assessments: Where we are and what we need. *Teachers College Press*.
- Doherty, J., & Conolly, M. (1985). How accurately can primary school teachers predict the scores of their pupils in standardised tests of attainment? A study of some non-cognitive factors that influence specific judgments. *Educational Studies*, 11(1), 41-60.
- Egan, O., & Archer, P. (1985). The accuracy of teachers' ratings of ability: A regression model. *American Educational Research Journal*, 22(1), 25-34.
- Elmore, R. F. (2004). *School reform from the inside out: Policy, practice, and performance*. Cambridge, MA: Harvard Education Press.
- Gardner, E. (1982). *Ability testing: Uses, Consequences, and Controversies*. Washington, D.C.: National Academy Press.

- Gilman, D. A., & Reynolds, L. L. (1991). The side effects of state wide testing. *Contemporary Education*, 62(4), 273-278.
- Ginsbery, M.B., & Murphy, D. (2002). How walkthroughs open doors. *Educational Leadership*, 59(8), 34-36.
- Gipps, C., & Murphy, P. (1994). A fair test?: Assessment, achievement and equity. Buckingham, UK: Open University Press.
- Glaser, R., & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. *Review of Research in Education*, 20, 393-419.
- Green, K. E., & Stager, S. F. (1986). Measuring attitudes of teachers toward testing. *Measurement and evaluation in counseling and development*, 19(3), 141-150.
- Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.
- Guthrie, J. W., & Springer, M. G. (2004). *A Nation at Risk* revisited: Did “wrong” reasoning result in “right” results? At what cost? *Peabody Journal of Education*, 79(1), 7-35.
- Haladyna, T., Haas, N., & Allison, J. (1998). Continuing tensions in standardized testing. *Childhood Education*, 74(5), 262-273.
- Helmke, A., & Schrader, F. W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education*, 3(2), 91-98.
- Herman, J. L., Abedi, J., & Golan, S. (1994). Assessing the effects of standardized testing on schools. *Educational and Psychological Measurement*, 54(2), 471-482.
- Hiebert, J. (2003). What research says about the NCTM standards. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

- Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology*, 76(5), 777-781.
- Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, 59(3), 297-313.
- Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan*, 81, 199-203.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA.
- Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practice*, 27(2), 28-45.
- Lawrenz, F., Huffman, D., & Appeldoorn, K. (2002). *Classroom observation handbook*. Minneapolis, MN: The College of Education and Human Development, University of Minnesota.
- Leinhardt, G. (1983). *Novice and expert knowledge of individual student's achievement*. Pittsburgh, PA: Learning Research and Development Center.
- Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, 32(7), 3-13.
- Lipman, P. (2004). *High stakes education: Inequality, globalization, and urban school reform*. New York: RoutledgeFalmer.
- Macpherson, R. J. S. (1998). Contractual or responsive accountability? Neo-centralist 'self-management' or systemic subsidiarity! Tasmanian parents' and other stakeholders' policy preferences. *Australian Journal of Education*, 42(1), 66-89.

- Marshall, C., & Rossman, G. B. (2006). *Designing qualitative research* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.
- McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.
- Mehrens, W. A. (2002). Consequences of assessment: What is the evidence. In G. Tindal, & T. M. Haladyna (Eds.), *Large-scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Messner, J. (1952). *Social ethics: Natural law in the modern world*. St. Louis, MO: B. Herder Book Co.
- Miller, S. A., & Davis, T. L. (1992). Beliefs about children: A comparative study of mothers, teachers, peers, and self. *Child Development*, 63(5), 1251-1265.
- Miller, S. A., Manhal, M., & Mee, L. L. (1991). Parental beliefs, parental accuracy, and children's cognitive performance: A search for causal relations. *Developmental Psychology*, 27(2), 267-276.
- Moore, J. L., & Waltman, K. (2007, April). *Pressure to increase test scores in reaction to NCLB: An investigation of related factors*. Paper presented at the meeting of the American Educational Researchers Association, Chicago, IL.
- National Research Council (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, & B. Findell (Eds.), Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Noddings, N. (2007). *When school reform goes wrong*. New York: Teachers College Press.

Noddings, N. (2005). *The challenge to care in schools: An alternative approach to education* (2nd ed.). New York: Teachers College Press.

Northwest Evaluation Association (2005, August). *NWEA instructional resources: 10 ways to use the class breakdown by overall RIT and class breakdown by goal reports in the classroom*. Retrieved November 29, 2006, from http://www.nwea.org/assets/documentLibrary/10_Ways_Goal.pdf

Northwest Evaluation Association (2006). *NWEA Members*. Retrieved November 29, 2006, from www.nwea.org.

Northwest Evaluation Association (2006, July). *Step 1 MAP administration: Teachers and staff*. Retrieved November 29, 2006, from <http://www.nwea.org/assets/documentLibrary/Step%201%20%20Teacher%20PPT.pdf>

Northwest Evaluation Association (2007, May). *Annotated MAP reports*. Retrieved February 1, 2009 from <http://www.nwea.org>

Northwest Evaluation Association (2008). *The RIT scale*. Retrieved February 1, 2009 from <http://www.nwea.org>

Northwest Evaluation Association (2008, August). *Understanding teacher reports*. Retrieved February 1, 2009 from <http://www.nwea.org>

Northwest Evaluation Association (2008, September). *RIT scale norms: For use with Measures of Academic Progress*. Retrieved February 1, 2009 from <http://www.nwea.org>

Northwest Evaluation Association (2009, January). *Teacher handbook: Measures of Academic Progress (MAP)*. Retrieved February 1, 2009 from <http://www.nwea.org>

Northwest Evaluation Association (2009, July). *NWEA's Measures of Academic Progress is selected as a state-approved formative assessment in Colorado*. Retrieved December 1, 2009 from <http://www.nea.org/about-nwea/news-and-event/nweas-measures-academic-progress-selected-state-approved-formative-assess>

Northwest Evaluation Association (2009, November). *A study of the alignment of the NWEA RIT scale with the Kansas Assessment System*. Retrieved March 22, 2010 from <http://www.nwea.org/sites/www.nwea.org/files/resources/Kansas%20Alignment%20Report%20110709.pdf>

Northwest Evaluation Association (2010). *DesCartes: Transforming results into instruction*. Retrieved June 21, 2010 from <http://www.nwea.org/products-services/classroom-resources/descartes>

Parker High School (2009). PHS Walkthrough Observation Feedback Form. Retrieved April 1, 2009 from <http://www.parkerusd.org/phs/>

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy.

Pezdek, K., Berry, T., Renno, P. A. (2002). Children's mathematics achievement: The role of parents' perceptions of their involvement in homework. *Journal of Educational Psychology*, 94(4), 771-777.

Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

- Porter, A., & Chester, M. (2002). Building a high-quality assessment and accountability program: The Philadelphia example. *Brookings Papers on Education Policy*, 285-337.
- Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.
- Richardson, J. (2001). Seeing through new eyes: Walk throughs offer new way to view schools. *Tools for Schools*
- Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right*. Washington, DC: Economic Policy Institute.
- School District Reference. (2006, May 18). District adopts NWEA-MAP test. *Staff notebook: A newsletter for all employees*, 36, 1-2. Retrieved November 29, 2006, from internal district website.
- School District Reference. (2006, July 14). *Suggested talking points for how NWEA MAP testing complements and informs [the state] formative assessments*. Retrieved November 29, 2006, from internal district website.
- School District Reference. (2006, July 25). *MAP input meeting notes from high, middle, and elementary school sessions*. Retrieved November 29, 2006, from internal district website.
- School District Reference. (2006, December 4). *Puttin' on the RITS*. Retrieved December 12, 2006, from school district internal website.
- Schumacher, E. F. (1973). *Small is beautiful: Economics as if people mattered*. New York: Harper & Row, Publishers, Inc.
- Sever, D. (2004). *Dancing with data to improve learning*. Lanham, MD: Scarecrow Education.
- Skretta, J. & Fisher, V. (2002). The walk-through crew. *Principal Leadership*, 3(3), 39-41.

- Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, 20(5), 8-11.
- Stiggins (2004). New assessment belief for a new school mission. *Phi Delta Kappan*, 88(1), 22-27.
- Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, 13(1), 12-28.
- Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. *Review of Research in Education*, 9, 377-435.
- Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, 68(2), 202-248.
- Tuhiwai Smith, L. (1999). *Decolonizing methodologies: Research and indigenous peoples*. New York: St. Martin's Press.
- Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.
- Wantanabe, M. (2007). Displaced teacher and state priorities in a high-stakes accountability context. *Educational Policy*, 21(2), 311-368.
- Wheelock, A., Bebell, D. J., & Haney, W. (2000). What can student drawings tell us about high-stakes testing in Massachusetts? *Teachers College Record*. Retrieved November 19, 2009 from, <http://www.tcrecord.org>
- William, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester, Jr. (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning*. Charlotte, NC: Information Age Publishing.

Wilson, L. D. (2007). High-stakes testing in mathematics. In F. K. Lester, Jr. (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning*. Charlotte, NC: Information Age Publishing.

Wright, D., & Wiese, M. J. (2001). Teacher judgment in student evaluation: A comparison of grading methods. *Journal of Educational Research*, 82(1), 10-14.

Appendix A: Semi-Structured Administrator Interview Protocol

I. Purposes and Effects of NWEA-MAP Testing?

1. Why does the district use the NWEA-MAP assessment?
 - a. Do you have any insight as to why they chose this particular test?
2. Do you think this test is meeting the district's purposes?
 - a. For your students? Your teachers? Your school? The district as a whole?
3. What are the costs of using this test?
 - a. Financial? Time? Staff? Other?
 - b. How much time is spent on the use of this test? On testing, in general?

II. What's Changed?

4. If you think back to what teaching at this school was like several years ago -- prior to the implementation of the NWEA-MAP test -- and compare that to now, what has changed?
 - a. Are there other initiatives that have been implemented during that time?
 - b. Are there other factors that might be related to those changes?
5. How do you use the NWEA-MAP data?
 - a. Do you feel prepared to interpret the results?
 - b. Are any decisions based primarily or in large part on these scores?
6. Is this test having any other effects -- positive or negative?

III. Reactions?

7. How do students react to the NWEA-MAP assessment?
8. How do teachers react to the NWEA-MAP assessment?

IV. Overall View of NWEA-MAP?

9. If you had a choice, would you want your school to use this test? Why or why not?
 - a. Do you have any concerns about the way the test is administered?
 - b. What changes could be made to improve the use of this test?

Conclusion of Interview: Do you have any other comments that you would like to share about the NWEA-MAP assessment?

Appendix B: Semi-Structured Teacher Interview Protocol

I. Purposes and Effects of NWEA-MAP Testing?

1. What do you think are the district's purposes for using the NWEA-MAP assessment?
 - a. Why do you think this? Based on meetings? Documents? General perception?
2. Do you think this test is meeting the district's purposes?
 - a. For you? Your students? Your school? The district as a whole?
3. Is this test having any other effects – positive or negative?

II. What's Changed?

4. If you think back to what teaching at this school was like several years ago -- prior to the implementation of the NWEA-MAP test -- and compare that to now, what has changed?
 - a. Are there other initiatives that have been implemented during that time?
 - b. Are there other factors that might be related to those changes?
5. After the last administration of the NWEA-MAP test in December, did you change anything with your students?

III. Useful for Teachers?

6. How does the NWEA-MAP assessment affect the way that you teach math?
 - a. Any effect on teaching methods and strategies?
 - b. Any effect on the content of your mathematics instruction?
7. How do you use the NWEA-MAP data?
 - a. Do you feel prepared to interpret the results?
 - a. Do you discuss the results with other teachers?
8. Do some teachers collaborate more than others?
 - a. If so, why? What are the characteristics of these teachers?

IV. Student Reactions?

9. How do students react to the NWEA-MAP assessment?
 - a. Do they view it differently than other assessments that they have to take?
 - b. Do they take it seriously? Get nervous? Want to know their scores?

V. Overall View of NWEA-MAP?

10. How do you feel about standardized testing, in general?
11. If you had a choice, would you want your students to take the NWEA-MAP assessment? Why or why not?
 - a. Do you have any concerns about the way the test is administered?
12. Is there a way that this test could be more useful to you as a teacher?

Conclusion of Interview: Do you have any other comments that you would like to share about the NWEA-MAP assessment?

Appendix C: Teacher Survey

Teacher Survey: The NWEA-MAP Mathematics Assessment

Thank you for making time to complete this survey. You do not need to provide your name anywhere on this survey. Your completed survey will not be shown to anyone outside of the University of Kansas. Please choose only one answer for each question (unless otherwise indicated) and use the last page for any comments or clarifications.

Demographic Info:

How many years have you been a teacher? _____

How many years have you been teaching at this school? _____

Please circle your age range:

under 25 25-34 35-44 45-54 55-64 65+

From which college/university did you obtain your bachelor's degree? _____

What year did you complete your bachelor's degree? _____

If you have obtained any subsequent degrees related to teaching/education, what year(s) did you complete these? _____

In what areas are you licensed to teach in the State of Kansas? _____

Please Indicate Your Level of Agreement with Each Statement by Circling Your Response:

SA
Strongly
Agree

A
Agree

D
Disagree

SD
Strongly
Disagree

Please ONLY consider the MATHEMATICS portion of the NWEA-MAP assessment.

- | | | | | |
|--|----|---|---|----|
| 1. The NWEA-MAP assessment is <i>as accurate</i> a measure of student achievement as a teacher's judgment. | SA | A | D | SD |
| 2. The NWEA-MAP assessment is compatible with my district's mathematics curriculum. | SA | A | D | SD |
| 3. Overall, the benefits of the NWEA-MAP assessment are worth the time invested. | SA | A | D | SD |
| 4. The instructional texts and materials that the district requires me to use are compatible with the NWEA-MAP assessment. | SA | A | D | SD |
| 5. Scores on the NWEA-MAP assessment accurately reflect the <i>quality of education</i> students have received. | SA | A | D | SD |
| 6. The NWEA-MAP assessment is just another fad. | SA | A | D | SD |
| 7. NWEA-MAP assessment is NOT an accurate measure of what my students know and can do. | SA | A | D | SD |
| 8. The majority of my students try their best on the NWEA-MAP assessment. | SA | A | D | SD |
| 9. Many students are extremely anxious about taking the NWEA-MAP assessment. | SA | A | D | SD |
| 10. The NWEA-MAP assessment has brought much needed attention to education issues in my school. | SA | A | D | SD |
| 11. Administrators in my school believe students' NWEA-MAP assessment scores reflect the quality of teachers' instruction. | SA | A | D | SD |
| 12. Professional development regarding the NWEA-MAP assessment has adequately prepared me to <i>administer</i> the test. | SA | A | D | SD |

13. Professional development regarding the NWEA-MAP assessment has adequately prepared me to <i>interpret</i> the test results.	SA	A	D	SD
14. Professional development regarding the NWEA-MAP assessment has adequately prepared me to <i>apply</i> the results to my teaching.	SA	A	D	SD
15. Taking the NWEA-MAP assessment is a good use of student time.	SA	A	D	SD
16. The NWEA-MAP assessment provides me with new information about my students that I would not otherwise know.	SA	A	D	SD
17. The NWEA-MAP score reports are easy to interpret.	SA	A	D	SD
18. The NWEA-MAP score reports provide useful information.	SA	A	D	SD

19. Do you specifically use the results of the NWEA-MAP assessment for any of the following activities? **[Check All That Apply]**

- | | |
|---|--|
| <input type="checkbox"/> Group students within my class | <input type="checkbox"/> Plan curriculum |
| <input type="checkbox"/> Evaluate student progress | <input type="checkbox"/> Give feedback to students |
| <input type="checkbox"/> Assess my teaching effectiveness | <input type="checkbox"/> Give feedback to parents |
| <input type="checkbox"/> Select instructional materials | <input type="checkbox"/> Determine student grades |
| <input type="checkbox"/> Plan my instruction | <input type="checkbox"/> Plan for remediation for students |
| <input type="checkbox"/> None of the above | |

20. To what extent do you feel pressured by each of the following to increase your students' scores on the NWEA-MAP assessment? **[Circle the appropriate word/phrase on each line]**

a. <i>Yourself</i>	a lot	a little	none
b. <i>Colleagues</i>	a lot	a little	none
c. <i>School-level Administrators</i>	a lot	a little	none
d. <i>School District</i>	a lot	a little	none
e. <i>Parents</i>	a lot	a little	none

As you know, the NWEA-MAP assessment results are reported in terms of RIT (Rasch Unit) Scores. Please indicate whether the following statements about the NWEA-MAP assessment and RIT scores are *true* or *false* by circling your response:

- | | | | |
|--|------|----|-------|
| 1. The RIT Scale helps to measure growth over time. | True | or | False |
| 2. The RIT Scale has the same meaning regardless of grade or age of the student. | True | or | False |
| 3. When making a placement decision for a student, the single RIT score (rather than the score range) should be used to help make the determination. | True | or | False |
| 4. NWEA-MAP assessments differ for each grade-level. | True | or | False |
| 5. If a student's RIT score increases significantly, then his/her percentile score should also show a significant increase. | True | or | False |
| 6. If a student took the NWEA-MAP test twice in one week, the student's score would almost always (in at least 90% of cases) fall within the same RIT range. | True | or | False |
| 7. If the score report indicates that a student's percentile rank is 65%, this means that the student scored better than 65% of the students in the class. | True | or | False |
| 8. The student in the class with the highest RIT score probably answered only about half of the questions correctly. | True | or | False |
| 9. It is reasonable to expect an increase of 10 points in the RIT score of every student from the beginning to the end of the year. | True | or | False |
| 10. RIT scores range from about 100 to 300. | True | or | False |

Please Indicate Your Level of Agreement with Each Statement by Circling Your Response:

	SA Strongly Agree	A Agree	D Disagree	SD Strongly Disagree
1. When a student does better than usual in mathematics, it is often because the teacher exerted a little extra effort.	SA	A	D	SD
2. I continually find better ways to teach mathematics.	SA	A	D	SD
3. Even when I try very hard, I do not teach mathematics as well as I teach most subjects.	SA	A	D	SD
4. When the mathematics grades of students improve, it is often due to their teacher having found a more effective teaching approach.	SA	A	D	SD
5. I know how to teach mathematics concepts effectively.	SA	A	D	SD
6. I am not very effective in monitoring mathematics activities.	SA	A	D	SD
7. If students are underachieving in mathematics, it is most likely due to ineffective mathematics teaching.	SA	A	D	SD
8. I generally teach mathematics ineffectively.	SA	A	D	SD
9. The inadequacy of a student's mathematics background can be overcome by good teaching.	SA	A	D	SD
10. When a low-achieving child progresses in mathematics, it is usually due to extra attention given by the teacher.	SA	A	D	SD
11. I understand mathematics concepts well enough to be effective in teaching elementary mathematics.	SA	A	D	SD
12. The teacher is generally responsible for the achievement of students in mathematics.	SA	A	D	SD
13. Students' achievement in mathematics is directly related to their teacher's effectiveness in mathematics teaching.	SA	A	D	SD

14. If parents comment that their child is showing more interest in mathematics at school, it is probably due to the performance of the child's teacher.	SA	A	D	SD
15. I find it difficult to use manipulatives to explain to students why mathematics works.	SA	A	D	SD
16. I am typically able to answer students' questions.	SA	A	D	SD
17. I am unsure if I have the necessary skills to teach mathematics.	SA	A	D	SD
18. Given a choice, I do not invite the principal to evaluate my mathematics teaching.	SA	A	D	SD
19. When a student has difficulty understanding a mathematics concept, I usually am at a loss as to how to help the student understand it better.	SA	A	D	SD
20. When teaching mathematics, I usually welcome student questions.	SA	A	D	SD
21. I do not know what to do to turn students on to mathematics.	SA	A	D	SD
22. Teacher morale is high in my school.	SA	A	D	SD
23. Teachers have high expectations for the in-class academic performance of students in my school.	SA	A	D	SD
24. My school has an atmosphere conducive to learning.	SA	A	D	SD
25. My school is more interested in increasing test scores than in improving overall student learning.	SA	A	D	SD

Use this Space to Make any Additional Comments.

If you are responding to a particular item on this survey, please indicate the page number and item number along with your response.

Appendix D: Pre-MAP Testing Lesson Observation Form

CETP Core Evaluation – Classroom Observation Protocol

CETP – CORE EVALUATION CLASSROOM OBSERVATION PROTOCOL

I. Background Information

A. Observer

1. Name: _____
2. CETP: _____ Institution Name: _____
3. Date of Observation: _____
4. Length of observation: _____ (minutes)
5. Was the teacher informed about this observation prior to the visit? ☐ Yes ☐ No

B. Teacher/Faculty

1. Name: _____
2. CETP Teacher? ☐ Yes ☐ No
3. Gender: ☐ Male ☐ Female
4. K-12: Licensure/certification _____
OR College Rank: (Check one.)
☐ Instructor/Adjunct Faculty ☐ Full Professor
☐ Assistant Professor ☐ TA: primary responsibility? _____
☐ Associate Professor ☐ Other: _____

II. Classroom Demographics

- A. What is the total number of students in the class at the time of the observation?
☐ 15 or fewer ☐ 26–30 ☐ 61–100
☐ 16–20 ☐ 31–40 ☐ 101 or more
☐ 21–25 ☐ 41–60

- B. Was a paraprofessional or teaching assistant in the class?
☐ Yes ☐ No

- C. 1. Grade Level (K-12) _____

OR

2. Student Audience (majority of students. Check any that apply):
 (a) ☐ Prospective teachers: (1) ☐ Elementary (2) ☐ M.S. (3) ☐ H.S.
 (b) ☐ Liberal Arts Majors
 (c) ☐ Mathematics/Science Majors

- D. Subject Observed/Descriptive Course Title: _____

- E. Scheduled length of class _____ (minutes)

B. In a few sentences, describe the lesson you observed and its purpose.

Include where this lesson fits in the overall unit of study, syllabus, or instructional cycle.

Note: This information needs to be obtained from the teacher/faculty member.

V. Ratings of Key Indicators

In this section, you are asked to rate each of a number of key indicators as descriptive of the lesson in five different categories, from 1 (not at all) to 5 (to a great extent). Note that any one lesson may not provide evidence for every single indicator; use DK, “Don’t Know,” when there is not enough evidence for you to make a judgment. Use N/A, “Not Applicable,” when you consider the indicator inappropriate given the purpose and context of the lesson.

1. This lesson encouraged students to seek and value alternative modes of investigation or of problem solving.	1	2	3	4	5	DK	N/A
2. Elements of abstraction (i.e., symbolic representations, theory building) were encouraged when it was important to do so.	1	2	3	4	5	DK	N/A
3. Students were reflective about their learning.	1	2	3	4	5	DK	N/A
4. The instructional strategies and activities respected students’ prior knowledge and the preconceptions inherent therein.	1	2	3	4	5	DK	N/A
5. Interactions reflected collaborative working relationships among students (e.g., students worked together, talked with each other about the lesson), and between teacher/faculty member and students.	1	2	3	4	5	DK	N/A
6. The lesson promoted strongly coherent conceptual understanding.	1	2	3	4	5	DK	N/A
7. Students were encouraged to generate conjectures, alternative solution strategies, and ways of interpreting evidence.	1	2	3	4	5	DK	N/A
8. The teacher/faculty member displayed an understanding of mathematics/science concepts (e.g., in her/his dialogue with students).	1	2	3	4	5	DK	N/A
9. Appropriate connections were made to other areas of mathematics/science, to other disciplines, and/or to real-world contexts, social issues, and global concerns.	1	2	3	4	5	DK	N/A

For the following questions, select the response that best describes your overall assessment of the *likely effect* of this lesson in each of the following areas, from 1 (no effect) to 5 (great effect).

10. Students’ understanding of mathematics/science as a dynamic body of knowledge generated and enriched by investigation	1	2	3	4	5	DK	N/A
11. Students’ understanding of important mathematics/science concepts	1	2	3	4	5	DK	N/A
12. Students’ capacity to carry out their own inquiries	1	2	3	4	5	DK	N/A

From: Lawrenz, F., Huffman, D., & Appeldoorn, K. (2002). *Classroom observation handbook*. Minneapolis, MN: The College of Education and Human Development, University of Minnesota.

Appendix E: Post-MAP Testing Classroom Observation Form

PHS Walkthrough Observation Feedback Form

Teacher Name	Teaching Methods						Cognition Level				Use of Standards				Student Engagement				
	1. SA EL	2. SLC	3. TLI	4. SW/TE	5. SW/TE	6. TD	KU	A	C	KR	CO/A	CO/NA	PO	UO	NO	100-81%	60-41%	40-21%	20-0%
Total for PHS																			

Keys

Section A: Teaching Methods (Painter & Valentine, 2002). Check each practice observed.

- Student Active Engaged Learning** (project work, cooperative learning, hands-on, demonstrations, active research, *Higher order thinking evident*.)
- Student Learning Conversations** (active conversation with all nearly all students engaged, all relevant ideas are encouraged and discussed (divergent thinking), teacher initiated, but not directed. *Higher order thinking*)
- Teacher-Led Instruction** (lecture, question and answer, teacher giving instructions, video instruction with teacher interaction, discussion may occur but instruction and ideas come primarily from teacher)
- Student Work/Teacher Engaged** (students doing seatwork, worksheets, book work, tests, individual reading, independent or group work, with teacher providing assistance to individuals or groups of students)
- Student Work/Teacher Disengaged** (students doing seatwork, worksheets, book work, tests, individual reading, independent or group work, with teacher doing something not related to the learning tasks of students)
- Total Disengagement** (students and teacher not engaged in activities associated with learning curriculum)

Section B: Level of Cognition (R. Marzano's Taxonomy) Circle one- the level most frequently observed.

Knowledge Utilization	Analysis	Comprehension	Knowledge Retrieval
<i>Apply or use knowledge in a new or specific situation:</i> Solve a problem, make decision, conduct an experiment, investigate, produce a product, judge, evaluate, resolve, create, plan, design, compose, make up	<i>Examine knowledge in fine detail and, as a result, generate new conclusions:</i> Classify, identify errors in thinking, generalize, specify, match, defend, compare/contrast, distinguish fact/opinion, construct & defend new conclusions	<i>Identify the key elements of information—get the essential meaning:</i> Summarize info, condense meaning, main idea, express in a graph or other non-linguistic representation	<i>Recall or execution of knowledge as previously learned:</i> Define, remember, who, what, where, when, how, describe, choose, how much, what part doesn't fit, pick an example, show, practice a skill (i.e. role math problems, a song, physical activities, etc.)

Section C: Use of Standards/Objectives Circle one- the level most frequently observed.

Clear Objective - Aligned	Clear Objective-Not aligned	Process Objective	Unclear Objective	No Objective
Clear objective aligned to state standards at the appropriate grade level/level of difficulty	Clear objective but not aligned to state standards at appropriate grade level and/or level of difficulty	Necessary objective not found in standards (work in groups, class rules, use a microscope, safety issues, etc.)	Learning-related activities but without clear objective directly aligned to standards	Activities without relevance to state standards or learning objectives

Comments:

From: Parker High School (2009). PHS Walkthrough Observation Feedback Form.

Retrieved April 1, 2009 from <http://www.parkerusd.org/phs/>

Ad Maiorem Dei Gloriam